Measuring an orienteer's performance: a method based on lower rank approximation of asymmetric matrices

Thesis

Measuring an Orienteer's Performance: A method based on lower rank approximation of asymmetric matrices

by

Sofia Villers Gomez

BSc, MSc

A thesis submitted to The Open University

in fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics and Statistics

The Open University, UK

October 2014

Date of Submission: 22 October 2014
Date of Award: 10 February 2015

Abstract

The objective of this thesis is to develop a method that measures the performance of competitors in an orienteering course. The measures proposed will allow the comparison of the orienteer's performance with other competitors running the same course, and with his/her performance in different courses.

This thesis presents two main contributions, the first one refers to an orienteer's performance measure, and the second one relates to the lower rank approximation theory developed for this research. In the orienteering context, we developed a method that calculates a fitness measure and a navigational measure for each orienteer running a course. This differentiation between physical and navigational skills is an original approach to the orienteering performance analysis, and it allows a clear identification of areas that need improvement to achieve better times in an event. Based on the performance measures of orienteers participating in 100 events that took place in the UK between January 2013 and May 2014, we construct a course technical difficulty measure for green, blue and brown courses. This course difficulty allows the identification of easy and hard courses.

On the statistical theory context, this thesis presents a robust and asymmetric lower rank approximation algorithm to reduce matrices with asymmetric outliers into two vectors. This algorithm was based on the robust lower rank approximation algorithm proposed by Maronna and Yohai (2008). This thesis presents two original modifications to this algorithm, which produces better estimates when the data has asymmetric outliers. Those modifications are the definition of an asymmetric objective function, and the use of a robust

scale parameter that is updated at each iteration. The proposed modification to the scale parameter caused that the estimates do not depend anymore on initial values and the value reached by the loss function is most of the time better than the point reached without the modification. We also show that this methodology can be applied in contexts other than orienteering.

Acknowledgements

I would like to thank the examiners Prof. Philip Scraf, Prof. Frank Critchley and Prof. Paul Garthwaite for the time dedicated to this thesis, for the suggestions and comments to improve this work. A very special thank you to my supervisors Dr. Karen Vines and Prof. Kevin McConway for all their help, guidance and encouragement through this wonderful three years. To my parents and my sister for always being the source of love and support to accomplish my dreams.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Orienteering is an outdoor sport where participants are given a specially prepared orienteering map. The selected course is marked in the map with circles and numbers. The circles mark the location of the check points and the numbers the order in which they have to be covered. Using only this map and a compass the orienteers have to complete the designated course, choosing the fastest possible route and passing through all the check points. These courses are intended to test the physical and navigational abilities of the competitors.

The objective of this thesis is to develop a method that measures the performance of competitors in an orienteering course. The measures proposed will allow the comparison of an orienteer's performance with other competitors running the same course, and with his/her performance in different courses.

This thesis starts with a description of orienteering and a review of previous research on orienteer's performance in Chapter 2. From that review we highlight the need for a performance measure that considers both the navigational and fitness skills. This research is based on published data for orienteering events, which are

available online. These data correspond to the times given between check points for each orienteer on each course. Using the times from a particular event we study the characteristics of the orienteering data. This analysis will show that by the nature of the sport, orienteering data sets are skewed and have asymmetricly distributed outliers. Those outliers correspond to longer than usual times between check points, and usually these longer times are related to points along the course where the orienteers got lost. This means that those outliers might be informative about the orienteers' navigational skills, so we are interested in obtaining as much information as possible from the outliers at the same time of analysing the rest of the data. Considering these properties of the data, at the end of Chapter 2 we present a summary of the basic statistical knowledge on robust statistics and multivariate analysis.

The main theoretical proposal in this thesis is the use of a robust and asymmetric lower rank approximation algorithm to decompose the times done in a course by all the competitors into two vectors. One vector is associated with an orienteer's speed along the course and the other related to the distances between check points.

In Chapter 3 we present our algorithm which is based on three modifications of a robust lower rank approximation algorithm proposed by Maronna and Yohai (2008). These modifications are: first a different setting for the initial values; second the use of another robust estimator for the scale parameter that will be updated at each iteration; and third a stopping rule for the algorithm that only depends on the number of iterations. Also to improve the algorithm's performance for data with asymmetric outliers, we define an asymmetric function to be used as the loss function in the algorithm.

An analysis of our proposed algorithm is presented in Chapter 4. Through simulated data sets we analyse the performance of the proposed asymmetric objective function compared with other symmetric functions that are presented in Chapter 3. We also study the advantages of the modifications done to the algorithm proposed by Maronna and Yohai (2008), discuss the convergence of the algorithm to minimal values of the loss function, and analyse how sensitive the modified algorithm is to data sets that do not have clear asymmetric outliers.

As mentioned before, orienteering courses are intended to test both the physical and navigational abilities of orienteers. So based on the two vectors obtained from the robust and asymmetric lower rank approximation algorithm and on the residuals observed when modelling the data by these two vectors, in Chapter 5 we construct two fitness and three navigational measures.

To measure an orienteer's fitness level we use the average speed and the ranked average speed measures. The average speed measures how fast the orienteer was able to complete the course. On some courses it is inherently much easier to run faster compared with other courses, which means that the proposed speed measure is course dependent. So the average speed measure is a pseudo speed (in the sense that it does not only compare distance and time), because it also includes characteristics of the course. The dependency of the average speed measure on the course, makes the average speeds of an orienteer in two different courses not comparable. The ranked average speed is a modification of the average speed, we base the ranking on the fastest orienteers in that course. This ranked average speed measure allows the comparison of an orienteer's fitness across different courses.

An orienteer's navigational measures are based on the analysis of speed consistency, proportion of legs with mistakes and time lost in those mistakes. The three navigational measures together provide a clear view of an orienteer's navigational skills. The combined analysis of the two fitness and three navigational measures gives a more complete approach to the complex problem of modelling an orienteer's performance.

After analysing the performance for each orienteer, we go to the next level and study the courses. In Chapter 6 we focus on certain type of courses, known in orienteering as colour coded courses, that is courses which are labelled by colour. By definition the courses designed with the same colour should have a similar difficulty level. Our interest is to analyse the consistency of the difficulty level across same colour courses. To do that we compare green, blue and brown courses from different events in the UK. We analyse the times in each event, followed by studying the courses' technical difficulty based on the performance measures developed in Chapter 5. The chapter finishes with the proposal of a methodology to measure the technical difficulty of the courses based on the empirical distribution of 100 events that took place in the UK in 2013 and 2014.

In Chapter 7 we discuss how the methodology proposed in Chapter 3 can be applied to a problem that is not related to orienteering. We present the generalized characteristics of the orienteering problem, in particular focusing on the matrix that is being decomposed by the robust and asymmetric lower rank approximation algorithm. The analysis of how the data has to be in order for the method to be used, allows us to find other applications for the method. In particular we use the algorithm to study expenditure in Mexican households. With data from a 2010 Mexican

national survey we show that the method can be applied in this case to estimate the distribution of the households' expenditure, estimate missing values and detect possible outliers in the data. Results of applying the algorithm and the comparison of those results with published results are also presented in this chapter.

The contributions in this thesis are divided in two, the first refers to the orienteering application of the project, and the second relates to the statistical theory developed in the process of this research. In the orienteering application we have constructed two fitness and three navigational measures that together define an orienteer's performance in a course. Also we propose a method to measure the difficulty of courses based on the times registered after the events have occurred. On the statistical side we define an asymmetric objective function that produces better estimates if it is used in the robust lower rank approximation algorithm proposed by Maronna and Yohai (2008) when the data has asymmetric outliers. We also propose modifications of the initial values and scale parameter estimator of this algorithm leading to an improvement in the estimates when the lower rank approximating matrices are two vectors.

Chapter 2

Background and literature review

This chapter presents in Section 2.1 a description of orienteering and a review of previous research on orienteer performance. We will highlight the need for a performance measure that considers both the navigational and fitness skills. In Section 2.2 we will study the characteristics of orienteering data using the times from a particular event. This analysis will show that by the nature of the sport, orienteering data sets are skewed and have asymmetrically distributed outliers. Section 2.3 presents a summary of the basic statistical knowledge on robust statistics and multivariate analysis. This summary will attempt to cover the statistical background for more specific statistical method described in the following chapters.

2.1 Orienteering

Orienteering is a family of outdoor sports that involves racing against a clock and navigating on unfamiliar terrain through a number of points shown in a map. Among these sports, the oldest and the most popular is foot orienteering. Participants (called orienteers or runners) are given a topographical map, usually a specially prepared orienteering map. The objective is to complete a designated course, choosing the fastest possible route and passing through a number of check points in a prede-

termined order. The winner is the competitor completing the course in the shortest time.

The course design depends on the location of the event and the anticipated levels of skills and experience of the participants. The course is intended to test the navigation skill, concentration, and running ability of the competitors. The International Orienteering Federation (IOF) has set the competition rules for foot orienteering events. This rules give course planning guidelines for different competitions formats (sprint, middle distance, long distance, relay and sprint relay). Those guidelines specify the technical and physical difficulty level, the terrain and the winning time for each competition format. Based on the IOF rules in the UK the middle and long distance courses are mainly graded either by age and gender or by a colour code. In both cases guidelines for course design such as length and technical difficulty are given by the British Orienteering Federation (BOF).

We are interested in measuring the performance of competitors in an orienteering event. As the courses are intended to test the physical and navigational abilities, we suggest this measure should consider both the fitness and the navigational performance. An analysis of an orienteer's performance should be useful for detecting areas for improvement and offer a fair comparison between the performances of the orienteers on the course. Through the literature review carried out for this thesis, different studies about an orienteer's performance focusing in either the navigational or the physical skills were found. In the following paragraphs those previous research papers on orienteering performance will be discussed.

Scarf (2007) in his analysis of mountain marathons suggests there are three aspects to competing in an orienteering event: deciding which way to go, finding the check point on approach, and moving quickly through the terrain encountered on the route. The first two are part of the navigation skills where the route choice (deciding which way to go) involves identifying the potential routes and choosing the fastest among them. The third aspect depends on the running speed or fitness of the orienteer. Scarf (2007) also mentions that Naismith's rule is used by orienteers to choose the fastest route. This rule, usually used by walkers and runners, estimates how long it will take to complete a route which includes ascents. The basic walking rule suggest allowing 1 hour for every 5 kilometres, plus 1 hour for every 600 metres of ascent (Naismith, 1892).

There have been several studies about the relationship between the time for completing the course and the body physiological response (e.g. heart rate, oxygen uptake) of the orienteers during an orienteering event (Bird et al., 1993; Gjerset et al., 1997; Jensen et al., 1994, 1999; Johansson et al., 1988; Larsson et al., 2002; Moser et al., 1995; Peck, 1990). The analysis of performance during orienteering has been difficult to perform, since orienteers do not run the same route between check points (Larsson et al., 2002). To overcome this difficulty, Moser et al. (1995) and Gjerset et al. (1997) measured physiological variables on the orienteers as they passed through different check points of an orienteering event, and during a cross-country time-trial on a route previously run in an orienteering competition.

Johansson et al. (1988) suggest that running on a varied cross-country terrain, as occurs in orienteering, imposes a greater muscular load on the legs in comparison with other running events. Gjerset et al. (1997); Jensen et al. (1994, 1999); Larsson

et al. (2002); Moser et al. (1995) found a correlation between the time to complete an orienteering course and the maximal oxygen uptake. Larsson et al. (2002) also mention a negative association between mean relative heart rate and the amount of additional time spent on the course due to navigational mistakes, probably caused by the orienteers slowing down or even stopping to orientate themselves. These studies suggest that the physiological needs of orienteers are different to other running sports.

Larsson et al. (2002) suggest that besides endurance, the cognitive element in orienteering is another difference with other running events. The mental task of finding the best route between check points and avoiding mistakes is very important for being a successful orienteer. The works of Seiler (1996) and Ottosson (1996) focus on the cognitive demands during a competition, summarized under the following points:

- the selection of relevant map information for route choice
- the comparison between map and terrain in map reading
- the comparison between terrain and map in relocation
- the quick awareness of mistakes

These mental tasks are the navigational skill factor, but neither of these studies mention the effect of this cognitive process on an orienteer's performance.

Bird et al. (2001) analysed the changes in orienteering speeds related to differences between sexes and ages. Their analysis is based on the times taken by the three fastest men and three fastest women in each 5-year age groups in two national events (British Orienteering Championship and Jan Kelstrom trophy) in 1997 and 1999.

Their results show that women are 32% slower than men in orienteering terrain, but only 21% slower during cross-country races and 20% slower in laboratory tests. They also conclude that there is no difference in the speeds for orienteers aged between 21 and 40 year old, however between the ages 45 and 65 years old the orienteering speed of both men and women appears to linearly decline.

All the mentioned works analyse separately the importance of considering the fitness, navigation skills, age, gender and terrain factors when studying orienteering but they do not mention the effect on the competitor's performance of any combination between these factors.

Kolb et al. (1987) does feature such a combined approach. They developed a mathematical model for orienteering performance based on the results obtained by the Austrian elite-orienteers on a set of different tests designed to study certain abilities during their training. The model they proposed assumes that an orienteer's performance in an event depends on three factors (running ability, orienteering ability and running technique). They concluded that basic running ability and orienteering ability were the two main factors influencing orienteering performance, both of them contributing 46% to orienteering performance. The specific running technique in the terrain only contributed 8% to orienteering performance.

The research in this thesis intends to develop a model that measures the performance of the competitors in an orienteering event, after the event has taken place. This is in contrast to the model developed by Kolb et al. (1987). They used the results on different tests to predict expected performance (how the orienteer will do), our proposed model will use the resulting times after the event to analyse the

actual performance (how the orienteer did).

The factors that might influence an orienteer's performance can be grouped into two different types. The first type are orienteer factors. This group includes everything related to orienteer such as age, gender, fitness and navigation skills. The second type are course factors. This group includes environmental characteristics like terrain, season and weather; but also includes elements from the course design such as number of check points, length and metres of climb.

It is important to mention the existence of interaction between the factors. For example, the importance of navigation skills on performance depends on the terrain. Also the navigational skills might be more important on some course types compared to others. Another relevant fact to bear in mind is that from all the proposed factors the only ones that can be modified by the orienteer to obtain a better performance are fitness and navigation skills. This suggests that the analysis of orienteers' performance should be in terms of fitness and navigational performance. The objective of this research is to develop a method that measures the fitness and navigational performance based on the times that each orienteer took to go from one check point to the next one around the course.

2.2 Analysis of the data available

After each event, the times of all the competitors on the different courses are usually published. The club organizing the event publishes the times either on their own web page or on external pages such as WinSplits Online. (`http://obasen.orientering.se/winsplits/online/en/default.asp`)

The results are arranged as a matrix where each row represents an orienteer and the columns are the check points. So each matrix element is the time an orienteer takes to go from one check point to the next one. In orienteering terminology the stretch between two check points is called leg. The total length of the course is also known. This length corresponds to the shortest route a competitor could possibly take, independently of whether or not the orienteer follows that route. The technical difficulty of the course dictates generally how hard the navigation will be. The course technical difficulty in the published results is only identifiable by the relation between the course name and the BOF technical difficulty guidelines.

The course difficulty level, combined with the fitness and navigational ability of each orienteer, will have a direct impact in the time to complete the course, and in particular each leg. This time can be separated in two, on the one hand the time needed to go from one check point to the next one without getting lost, and on the other hand the time spent in any navigational mistake committed on that leg. Slower orienteers will have longer times in all the legs in comparison with the rest of the runners. But if an orienteer got lost in a leg the overall time to complete the leg will be longer in comparison to what was expected based on the times in the other legs, and this can occur to any orienteer in any leg. This means that the times recorded have a number of larger times distributed throughout all the matrix. If the larger time appears in one or few legs, it is sensible to assume that these times are related to orienteers getting lost in particular legs.

It is important to mention that:

- Sometimes orienteers do not complete the entire course.
- In some cases the technology used to record the times at a check point fails. This could occur for only one orienteer, a group or all the orienteers.

In both cases the data registered is either zero or blank. Because orienteers cannot take a zero amount of time to move from one check point to the next one, the zero is not a valid value in this context. So any zero or blank in the data can be treated as a missing value. These cases make the matrix of times incomplete. However the method that will be developed to measure an orienteer's performance allows a certain level of missing data. So we will include the times from orienteers from whom full data is not available.

To do a first analysis on the nature of data in an orienteering event we use the times published for a long course (Course 27) on the first day of the Scottish 6-day competition in 2013. This course had 14 legs, 143 participants and two of the orienteers had missing data on two legs. The next table shows the times for six of the orienteers. The names have been changed.

	leg 1 S-1	**leg 2** 1-2	**leg 3** 2-3	**leg 4** 3-4	**leg 5** 4-5	**leg 6** 5-6	**leg 7** 6-7	**leg 8** 7-8	**leg 9** 8-9	**leg 10** 9-10	**leg 11** 10-11	**leg 12** 11-12	**leg 13** 12-13	**leg 14** 13-F
Mary	1.18	0.78	1.80	1.67	2.78	1.20	2.65	2.38	11.33	1.00	3.38	3.72	1.07	0.43
Jackie	1.12	0.95	1.58	1.85	2.67	1.47	3.52	2.85	10.20	0.92	3.67	4.30	1.17	0.43
Sarah	1.50	1.27	2.05	4.90	3.72	1.45	3.63	3.88	12.63	1.35	5.32	20.18	1.57	0.52
Jane	2.82	5.83	1.82	2.52	3.27	1.88	6.33	2.78	11.00	2.67	5.43	25.48	1.70	0.47
Alice	4.22	4.00	4.32	5.88	4.17	3.37	8.82	5.62	27.18	5.03	8.93	7.77	2.05	0.50
Sam	3.32	2.78	4.98	5.58	14.07	3.42	9.00	5.22	28.05	7.62	12.87	18.72	2.17	0.58

Table 2.1: Example of the matrix of times for an orienteering course.

Here we see the time each orienteer takes to complete each leg, the scale is minutes on a decimal base, so Alice completed the second leg in exactly 4 minutes while Jackie did it in 57 seconds.

The next table presents the descriptive statistics of the times from the same course. The robust dispersion estimator is Sn, which was proposed by Rousseeuw and Croux (1993) as an alternative to the median absolute deviation that will be described in Section 2.3.1.

leg	mean	median	maximum	minimum	standard deviation	robust dispersion
leg 1	2.41	2.10	7.43	1.12	1.08	0.66
leg 2	1.87	1.52	11.95	0.78	1.42	0.48
leg 3	2.52	2.30	5.90	1.58	0.83	0.50
leg 4	3.66	3.08	10.57	1.67	1.76	0.97
leg 5	3.95	3.58	14.23	2.47	1.59	0.68
leg 6	2.09	1.92	6.63	1.20	0.70	0.46
leg 7	4.91	4.63	12.70	2.65	1.58	1.24
leg 8	3.81	3.62	8.63	2.38	0.93	0.73
leg 9	14.50	12.92	44.25	8.20	5.64	3.18
leg 10	2.60	1.97	17.20	0.92	2.01	0.82
leg 11	5.81	5.28	17.40	3.38	2.11	1.37
leg 12	8.52	7.40	25.48	3.72	3.84	2.62
leg 13	1.63	1.58	3.15	1.07	0.38	0.34
leg 14	0.48	0.47	1.10	0.32	0.10	0.07

Table 2.2: Descriptive statistics by leg.

Table 2.2 shows that the ninth leg was the one where orienteers spent most time on average completing, while the last leg, from the thirteenth control to the end of the course, was completed in the shortest mean time. The last leg also has the smallest dispersion which means that most of the competitors took a time very close to 0.48 minutes. The last leg is usually a short distance and there is no need to use any navigation technique. This is because the rules mention that the finish line must be clear to all competitors approaching it. So the small dispersion on the last leg is likely to be caused by the fact that navigational mistakes do not usually occur between the last check point and the finish.

It is important to mention the presence of asymmetry in the distribution of leg times. This is observed by noting that the difference between the mean and the minimum is smaller than the distance between the mean and the maximum. That asymmetry might be expected because the data are times and cannot have negative values, and can have very large positive values. In leg 3 and leg 10 we can see that legs with similar mean times have very different maximum times. The mean times for these

legs are 2.52 and 2.60 minutes respectively. However the maximum time of the third leg is 5.90 minutes compared with 17.20 minutes on the tenth leg. These suggest presence of at least one larger than usual time on leg 10. The large maximum times on other legs such as 44.25 minutes on leg 9 or 14.23 minutes on the fifth leg also suggest outliers in the data.

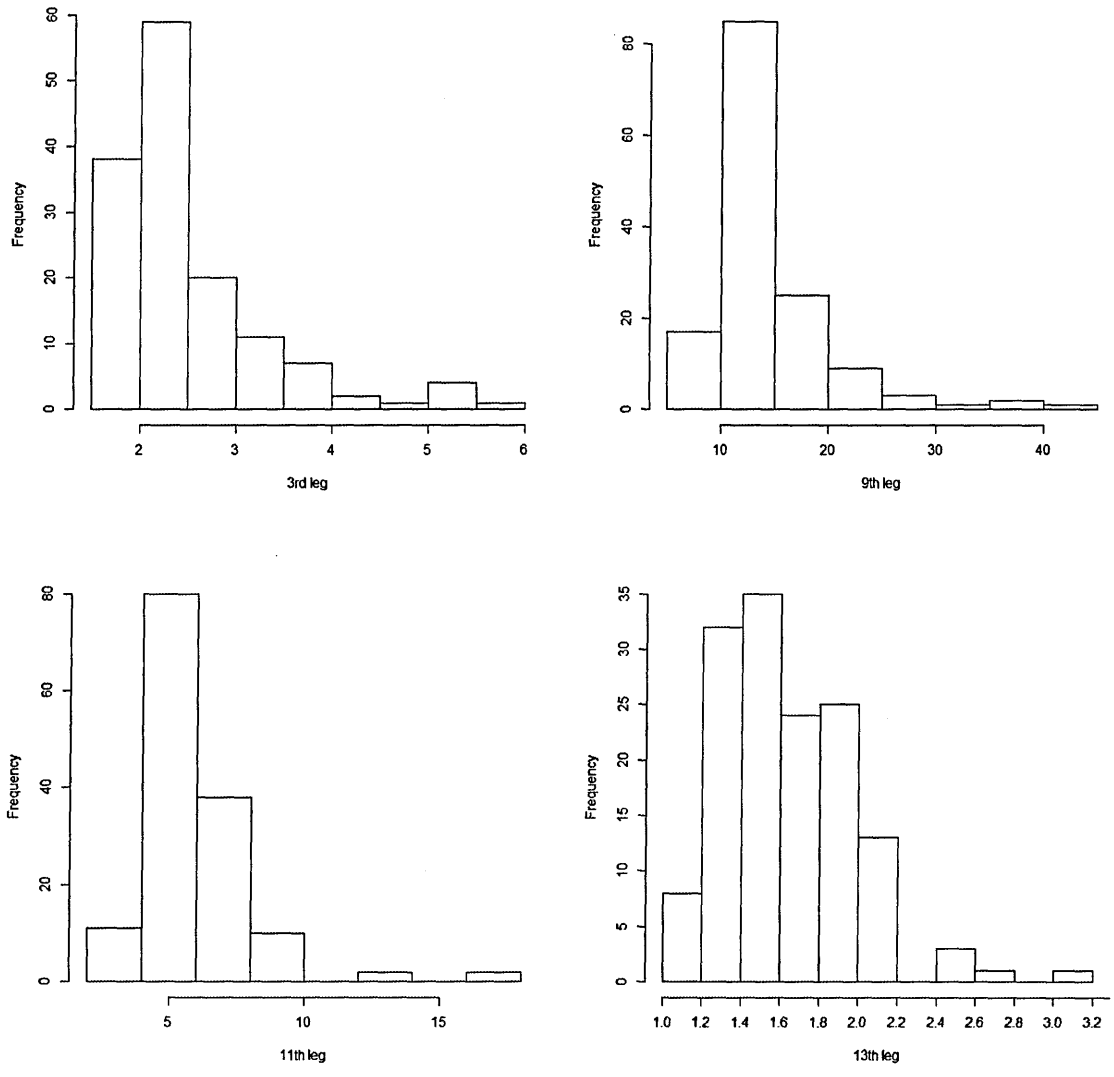
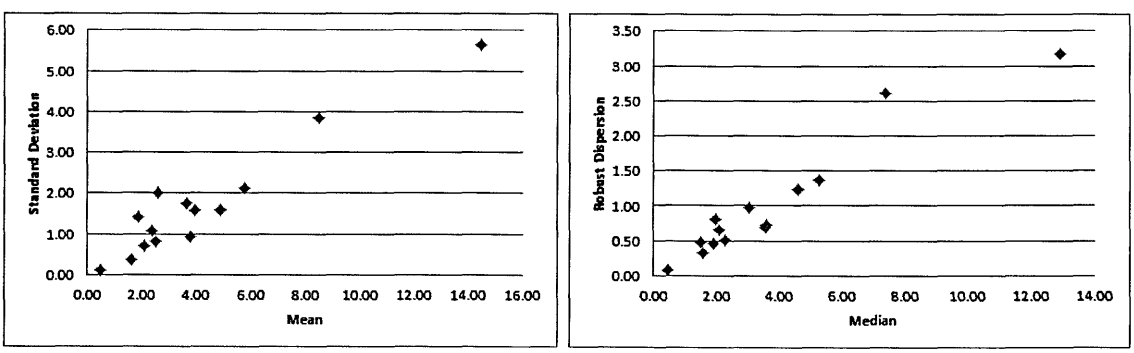


Figure 2.1: Histograms of the times on the third, ninth, eleventh and thirteenth legs

The histograms in Figure 2.1 show different examples of the distributions of the times per leg. We see that the distributions are positively skewed as we expected. This is because the data correspond to the times of all the orienteers to complete

a distance between two check points, and that data includes the times of orienteers that might have got lost in the that leg. This is the typical behaviour for leg times. The majority of the competitors completed leg 3 in a time between 1.5 and 3 minutes with 8 observations with values greater than 4 minutes. The ninth leg was completed by the majority of the orienteers in a time between 9 and 20 minutes with 4 larger times (greater than 30 minutes). The histogram for the eleventh leg shows the presence of outliers with some orienteers completing the leg in times over 10 minutes, which is almost twice the mean time for that leg. For the thirteenth leg all the competitors completed the leg in a time between 1.07 and 3.15 minutes, but the histogram shows a positively skewed distribution with 5 observations greater than 2.40 minutes. Also it can be seen that this leg in comparison with the other three shown in the figure has a smaller dispersion.

Figure 2.2 presents the relationship between the measure of central tendency and the measure of dispersion, based on the data for each leg. Figure 2.2a presents the plot of the mean and the standard deviation of the leg times. In Figure 2.2b the relationship between the robust statistics (median vs robust dispersion) is plotted.



(a) Scatter plot of the mean and standard deviation of the leg times.

(b) Scatter plot of the median and robust dispersion of the leg times.

Figure 2.2: Scatter plots of the descriptive statistics of times per leg.

Both Figures 2.2a and 2.2b show a positive linear relationship between the measures

of central tendency and the measures of dispersion. The relationship is stronger with the robust statistics as expected due to the presence of outliers in the data set. We conclude that for this set of orienteering data we observe a strong positive linear correlation between the median and the robust dispersion.

From this exploratory analysis and a general knowledge of orienteering, the following observations can be generalized to any course, and will help to have a better understanding of the orienteering data.

1. The leg lengths are different within one course and between courses. It also happens that in the same type of course the numbers of legs may be different.

2. The courses are not long enough to cause a decreasing effect on the orienteers' fitness. We assume that the orienteers' ability to run at a given speed any leg is the same independently of the order of the legs. However some orienteers might be much slower than the others.

3. Usually long distances are more demanding navigationally. An orienteer with good navigation skills will cover longer distances in less time than a competitor with worse navigation skills who needs to stop, set the map and find their way to the next point more frequently. This suggest that depending on the navigational skills of the orienteers running the course, longer legs might present more dispersion on the times to complete them.

4. The leg times do not have a symmetric distribution and some unusually large values are present. These large values could be caused by three reasons, first by orienteers who run slower than the majority of the competitors on the course, second particularly long legs and third by orienteers that got lost on that particular leg hence spending more time to complete the leg.

The histograms in Figure 2.1 and the statistics on Table 2.2 show the heterogeneity between the legs in one course. It was also shown that the orienteering data has outliers. The outliers are related to the navigational mistakes so they have a direct effect on the performance measure. So it is very important to analyse the outliers separately and assuming an asymmetric distribution is a sensible start. These characteristics of the data have to be considered in the statistical approach used to develop the performance measures.

2.3 Statistical background

As mentioned in the previous section the orienteering times have the characteristic of being data with a particular tendency to include outliers with noticeably high values. This suggests that the times per leg or per orienteer will not have symmetric distributions. As our interest is to analyse the performance of the orienteers through their times, we want to obtain as much information as possible from both the non outliers and the outliers in the data.

An outlier is a data point that deviates from usual assumptions and/or from the pattern suggested by the majority of the data. Outliers are more likely to occur in multivariate datasets. Because of the interaction between the variables it is often difficult to detect the outliers in multivariate data by simple visual inspection. Also when the data contains outliers, multivariate estimates differ substantially from the estimates that would be obtained without the outliers. However a robust fit makes the identification of the outliers possible (Hubert et al., 2008). For these reasons we propose the use of a robust statistical approach.

In this section we will describe the approaches studied in order to detect those orienteers for whom low fitness and/or navigational skills affected their performance. The methods proposed will be based on the idea that times related to low fitness and/or navigational skills will be different to the rest, so they can be seen as outliers. The section is divided in three parts, the first one gives the general basis of robust statistics, while the second and third present two approaches for the detection of outliers with their corresponding robust modification.

2.3.1 Robust statistics for univariate data

There exist robust parameter estimators that provide a good fit to the majority of the data when the data contain outliers, as well as when the data are free of them. A direct benefit of good fit to the majority of the data is the reliable detection of outliers, particularly in the case of multivariate data (Maronna et al., 2006).

The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is just the arithmetic average of the data. If any data value x_i varies from $-\infty$ to $+\infty$, the sample mean will change from $-\infty$ to $+\infty$ as well. This contrasts with the sample median, which is not much affected by the variation of any data value x_i (Maronna et al., 2006). We say that the median is resistant to gross errors whereas the mean is not. In fact the median is the most widely known robust estimator of the location parameter (Rousseeuw and Croux, 1993).

The situation for the scale parameter σ is similar. The classical estimator is the standard deviation (SD) and a single outlier can make SD arbitrarily large. A robust measure of scale is the median absolute deviation about the median (MAD),

defined as: for $x_1, ..., x_n$ a batch of numbers with median equal to $med(x)$

$$MAD(x) = med|x_i - med(x)| \tag{2.1}$$

To make the MAD comparable to the SD, the following modification is done:

$$MADN(x) = b\, med|x_i - med(x)| \tag{2.2}$$

where the constant b is needed to make the estimator consistent for the parameter of interest. In case of the usual parameter σ from a normal distribution $b = 1.4826$.

Rousseeuw and Croux (1993) mention that the MAD has the best possible breakdown point (50%). This means that this estimator will tolerate up to 50% gross errors before it can be arbitrarily large.

The extreme robustness of the sample median and the MAD make them ideal for screening data for outliers by computing for each observation x_i the statistic:

$$OUT(x_i) = \frac{|x_i - med(x)|}{MAD} \tag{2.3}$$

Using this statistic, Rousseeuw and Croux (1993) defined an outlier as an observation x_i for which $OUT(x_i)$ exceeds a certain cut-off (for example a cut-off of 2.5 or 3).

The MAD takes a symmetric view on dispersion, attaching equal importance to positive and negative deviations from the median. In other words the MAD finds the symmetric interval around the median and for this reason the use of this statistic seems natural in symmetric distributions but not in asymmetric ones (Rousseeuw

and Croux, 1993).

An alternative to the MAD suggested by Rousseeuw and Croux (1993) is the Sn statistic, defined as:

$$Sn = c\, med_i\{med_{j \neq i}|x_i - x_j|\} \tag{2.4}$$

In other words, for each i we compute the median of $\{|x_i - x_j|; j = 1, ..., n\}$. This yields n numbers, the median of which gives our final estimate Sn. The constant c is introduced for consistency similar to the b in the $MADN(x)$ equation, and its default value is $c = 1.1926$.

The Sn estimator differs from the MAD in the fact that it does not need any location estimate of the data. Instead of measuring how far away the observations are from a central value, Sn looks at a typical distance between observations, which make it not slanted towards symmetric distributions and more efficient (Rousseeuw and Croux, 1993).

2.3.2 Multivariate analysis

When more than one variable is observed for each individual, the end product of the data collection is a multivariate data set. Suppose $t_{i,j}$ denotes the observation corresponding to the jth variable as measured on the ith individual. The data are exhibited most simply as a data matrix, whose rows refer to the individuals in the sample and whose columns refer to the variables measured, denoted symbolically as follows: (Krzanowski, 1998).

$$\mathbf{T} = \begin{pmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,m} \\ t_{2,1} & t_{2,2} & \dots & t_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n,1} & t_{n,2} & \dots & t_{n,m} \end{pmatrix}$$

Multivariate analysis deals with situations in which several variables are measured simultaneously on each experimental unit, each being considered equally important at the start of the analysis. In most cases of interest it is known or assumed that some form of relationship exists among the variables, and hence that considering each of them separately would entail a loss of information (Maronna et al., 2006). Krzanowski (1998) states that the existence of the correlations between variables is central to multivariate analysis.

In the multivariate approach an outlier will refer to an observations of the ith individual in all the p-variables. This means that an outlier will be the row of the matrix $\mathbf{T}$ that are significantly different to the other rows of the matrix, called "atypical rows". The importance of the outliers in this approach is that their presences may alter the sample mean and/or covariances affecting the affine equivariance property assumed in some multivariate methods (Maronna et al., 2006).

Hubert et al. (2008) mention that because the usual multivariate analysis techniques (e.g., principal components, discriminant analysis and multivariate regression) are based on empirical means, covariance or correlation matrices, and least squares fitting, then they can be strongly affected by even a few outliers. They also comment on the possibility of the outliers affecting the estimated model so much "in their direction" that they will be well-fitted by it.

Rousseeuw and Van Zomeren (1990) comment on the classical method to detect outliers in the multivariate case. This method consists of computing the Mahalanobis distance for each point $\mathbf{t}_i = (t_{i,1}, t_{i,2}, \ldots, t_{i,m})$

$$MD_i = \sqrt{(\mathbf{t}_i - Z(\mathbf{T}))C(\mathbf{T})^{-1}(\mathbf{t}_i - Z(\mathbf{T}))^t} \tag{2.5}$$

where, $Z(\mathbf{T})$ is the arithmetic mean by column of the data set $\mathbf{T}$ as shown in the following equation:

$$Z(\mathbf{T}) = (\frac{1}{n}\sum_{i=1}^{n} t_{i,1}, \frac{1}{n}\sum_{i=1}^{n} t_{i,2}, ..., \frac{1}{n}\sum_{i=1}^{n} t_{i,m}) \tag{2.6}$$

and $C(\mathbf{T})$ is the usual sample covariance matrix.

The distance MD_i would give information about how far the vector $\mathbf{t}_i$ is from the centre of the cloud, taking into account the shape of the cloud as well. The cloud is formed by n vectors of the matrix $\mathbf{T}$, considering each vector as a point in a m-dimensional space. They stated that this approach suffers from the masking effect, by which multiple outliers do not necessarily have a large MD_i. The masking is caused by the fact that $Z(\mathbf{T})$ and $C(\mathbf{T})$ are not robust estimators, meaning that a small cluster of outliers will attract $Z(\mathbf{T})$ and will inflate $C(\mathbf{T})$ in its direction. Therefore, Rousseeuw and Van Zomeren (1990) suggest that the natural solution is to replace $Z(\mathbf{T})$ and $C(\mathbf{T})$ in (2.5) by robust estimators.

Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction. PCA reduces the number of variables to a small number of components that are linear combinations of the original variables. PCA uses the Euclidean distance to establish how far is one observation from the others. Because the Ma-

halanobis distance considers the covariance matrix in the calculations and the Euclidean distance does not, then the concept of outlier is different in the Mahalanobis distance and PCA methods (Maronna et al., 2006).

Like the Mahalanobis distance, an alternative approach in PCA is to replace the location and dispersion parameters by robust estimates. One option is to use M-estimates, however these estimates seem to depend on the shape of the dispersion estimate. Another procedure proposed in 1985 was to replace the variance used for the selection of the components by a robust scale estimator, this approach found serious computational problems (Maronna et al., 2006). Different robust approaches to PCA have been developed with their own advantages and disadvantages that have to be taken into account for each particular type of data.

It has been mentioned that the multivariate approach will identify "atypical rows" in the matrix. A disadvantage of this is that if the analysis is done for each column separately, it is possible that no outlier is detected, even though the row is defined as outlier. This means that this approach is unable to detect if one element of the matrix is significantly different to the other elements in the row and/or the column. This type of atypical elements are called element-wise outliers. In Section 2.2 it was mentioned that the orienteering data might present element-wise outliers (when the orienteer gets lost in a particular leg) and row-wise outliers or "atypical rows" (when very slow orienteers participate in the course). For those reasons any method used for the orienteering data should be able to detect element-wise and row-wise outliers.

2.3.3 Bilinear models

The reduction of high-dimensional data matrices into low-dimensional representations is a widely used multivariate data-analysis tool (Colin et al., 2008). The main idea of dimensional reduction is to approximate a matrix $\mathbf{T}$ of order $n \times m$, by another matrix $\mathbf{Y}$ also with dimensions $n \times m$ matrix but with rank p where $p < min(n, m)$. Fitting this matrix of rank p is equivalent to fitting a product matrix $\mathbf{AB}^t$ where the dimensions of $\mathbf{A}$ and $\mathbf{B}$ are $n \times p$ and $m \times p$ respectively. For orienteering data the use of $p = 1$ is a natural approach, where $\mathbf{A}$ is the orienteers' speed vector and $\mathbf{B}$ is the leg lengths vector.

Let the matrix $\mathbf{T}$ be a set of data of the form:

$$\mathbf{T} = \begin{pmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,m} \\ t_{2,1} & t_{2,2} & \dots & t_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n,1} & t_{n,2} & \dots & t_{n,m} \end{pmatrix}$$

$\mathbf{T}$ can be approximated by a bilinear model. The elements of the matrix will be fitted by:

$$t_{i,j} = \mathbf{a_i b_j}^t + e_{i,j} \tag{2.7}$$

where $\mathbf{a_i}$ and $\mathbf{b_j}$ denote columns of the matrices $\mathbf{A_{n\times p}}$ and $\mathbf{B_{m\times p}}$ respectively.

The classical least squares minimizing criterion for the approximation can be written as:

$$D_{LS}(\mathbf{T}, \mathbf{a}, \mathbf{b}) = \sum_{j=1}^{m}\sum_{i=1}^{n}(t_{i,j} - \mathbf{a_i b_j}^t)^2. \tag{2.8}$$

Householder and Young (1938) gave the exact solution of this least squares lower rank approximation. However, it is known that least squares estimators are sensitive to outliers, so the method mentioned is not suitable for data sets with outliers.

To achieve robustness Maronna and Yohai (2008) mention the existence of different approaches based on finding the directions that maximize or minimize the variability in the data, where the measure of variability is replaced by a robust dispersion estimate. This approach is resistant when a proportion of "atypical rows" is sufficiently small. However for situations where the contamination is not by row but instead each element could be randomly atypical, this approach fails.

Other robust estimates have been proposed by Croux et al. (2003) and Lui et al. (2003). Focusing on factor analysis models Croux et al. (2003) suggests the use of L_1 norm instead of L_2 norm. Lui et al. (2003) uses a least trimmed squares method to obtain a singular value decomposition of some microarray data. These approaches have the opposite problem, they are resistant to element-wise contamination but may fail when a row is completely atypical (Maronna and Yohai, 2008).

Another way to introduce robustness is by assigning weights to the elements of the matrix $\mathbf{X}$. Then the weighted least squares minimizing criterion becomes:

$$D_{WLS}(\mathbf{T}, \mathbf{a}, \mathbf{b}) = \sum_{j=1}^{m} \sum_{i=1}^{n} w_{i,j}(t_{i,j} - \mathbf{a_i b_j}^t)^2. \tag{2.9}$$

A direct solution for this weighted least squares problem does not exist. Gabriel and Zamir (1979) introduced an iterative weighted least squares algorithm to find a low-dimensional representation of a matrix $\mathbf{T}$ of $n \times m$. This algorithm assumes that the weights are known. In later work done by Verboon and Heiser (1994) they

suggest using iteratively reweighted least squares (IRLS) to estimate lower rank approximations when the weights are unknown (Colin et al., 2008).

Based on the ideas of the iterative weighted least squares algorithm proposed by Gabriel and Zamir (1979), Maronna and Yohai (2008) propose an alternating algorithm for low rank approximation matrices that produces estimates resistant to element-wise and row-wise contamination. This method is based on a column-scale estimate and M-estimators with a bounded ρ function (Maronna et al., 2006). We will describe the algorithm in more detail in Chapter 3.

It has been mentioned that orienteering data has asymmetric outliers, so we want the algorithm to be based on asymmetric M-estimator (AM-estimator). The literature shows that Allende et al. (2006) and Wang and Lee (2011) use asymmetric influence functions to obtain an AM-estimator. Motivated by the use of the Γ_A^0 distribution to model speckled imagery, which can be highly skewed, Allende et al. (2006) used a redescending asymmetric piecewise linear function to obtain the AM-estimator of the distribution parameters. Wang and Lee (2011) used an asymmetric bisquare objective function to obtain the AM-estimator of the Burr type III distribution parameters. In both papers the results show that the AM-estimator outperformed the maximum likelihood and traditional M-estimator methods. This means that defining an appropriate asymmetric ρ function in the algorithm might produce AM-estimators that perform better when asymmetric outliers are present.

Based on the alternating procedure proposed by Maronna and Yohai (2008) we will develop an algorithm to approximate a matrix with asymmetric outliers by a rank one matrix. The selection of a rank one approximation was based on the idea that as

the matrix correspond to times taken by orienteers running a series of legs, then the times can be estimated by the multiplication of the orienteers' speed vector and the leg lengths vector. So the rank one matrix will be in particular the multiplication of two vectors.

We will introduce an asymmetric function in the algorithm as an alternative to treat data sets with the asymmetric outliers. In the following two chapters we will explain step by step our proposed algorithm. We will show that it produces better estimates than the least squares or traditional M-estimators, with the advantage of having convergence in the weights and not needing the selection of starting points. Then in Chapter 5 we will use the results of this algorithm to estimate the orienteers' fitness and navigational performance using the times taken by the orienteers in a particular orienteering event.

Chapter 3

AM-estimator for lower rank approximation of matrices

In this chapter we will show the process we followed to go from the algorithm proposed by Maronna and Yohai (2008) to our modified algorithm mentioned at the end of the previous chapter and from now called the Improved M&Y algorithm. In Section 3.1 we explain in detail the theory behind robust lower rank approximation, focusing on the case of approximating an $n \times m$ data matrix by the multiplication of two vectors. Section 3.2 describes the construction of the asymmetric function which we propose should be used in the case of data with asymmetric outliers. Section 3.3 describes in detail each of the differences between our algorithm and the algorithm proposed by Maronna and Yohai (2008). At the end of Section 3.3 we present step by step the Improved M&Y algorithm that approximates a matrix of dimensions $n \times m$ by the multiplication of two vectors of dimensions n and m respectively.

3.1 Robust lower rank approximation of matrices

This section presents an outline of the robust lower rank approximation procedure proposed by Maronna and Yohai (2008) for matrices with atypical rows and scattered atypical elements. This method can be used to approximate an $n \times m$ data matrix with one of rank p, where the rank $p < min(n, m)$. As mentioned in Chapter 2 we are interested on the case of approximating a matrix $\mathbf{T}$ by rank one matrix and in particular when this rank one matrix comes from the multiplication of two vectors.

So the objective of the method is to reduce a matrix $\mathbf{T}$ of dimension $n \times m$, which might contain outliers, into two vectors $\mathbf{a}$ and $\mathbf{b}$, so that

$$\mathbf{T} = \begin{pmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,m} \\ t_{2,1} & t_{2,2} & \dots & t_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n,1} & t_{n,2} & \dots & t_{n,m} \end{pmatrix} \approx \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \dots & b_m \end{pmatrix} = \mathbf{a}\mathbf{b}^t$$

This means that each element of the matrix $\mathbf{T}$ will be fitted by a model of the form:

$$t_{i,j} = a_i b_j + error \tag{3.1}$$

To find good estimators of a_i and b_j we focus on the error. The error is defined as $r_{i,j} = t_{i,j} - a_i b_j$, Then the matrix of errors will be:

$$\mathbf{R} = \begin{pmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,m} \\ r_{2,1} & r_{2,2} & \dots & r_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \dots & r_{n,m} \end{pmatrix}$$

We assume that the errors are independent and identically distributed, with density function f_0. So the likelihood function of $\mathbf{R}$ is:

$$L(\mathbf{R}) = \prod_{j=1}^{m}\prod_{i=1}^{n} L(r_{i,j}) = \prod_{j=1}^{m}\prod_{i=1}^{n} f_0(t_{i,j} - a_i b_j) \tag{3.2}$$

If we use logarithms, then the maximum-likelihood estimates (MLE) of $\mathbf{a}$ and $\mathbf{b}$ will be those values that minimize the function

$$-\log(L(\mathbf{R})) = \sum_{j=1}^{m}\sum_{i=1}^{n} -\log f_0(r_{i,j}) = \sum_{j=1}^{m}\sum_{i=1}^{n} \rho(r_{i,j}). \tag{3.3}$$

where $\rho(x) = -\log f_0(x)$.

Robustness in the estimation procedure is achieved with an appropriate selection of the ρ function that is used instead of the logarithm of the density function ($\log f_0(x)$). Then the parameter estimators obtained by minimizing the loss function $l(\mathbf{R}) = \sum_{j=1}^{m}\sum_{i=1}^{n} \rho(r_{i,j})$ are called M-estimators.

Section 2.2.4 of Maronna et al. (2006) defines a ρ-function as:

- $\rho(x)$ is a non-decreasing function of $|x|$
- $\rho(0) = 0$
- $\rho(x)$ is increasing for $x > 0$ such that $\rho(x) < \rho(\infty)$
- If $\rho(\infty) = 1$ then ρ is bounded.

The M-estimators proposed by Maronna and Yohai (2008) are minimisers of expressions of the form:

$$l(\mathbf{a}, \mathbf{b}, \sigma) = \sum_{j=1}^{m} \sigma_j^2 \sum_{i=1}^{n} \rho\Big(\frac{r_{i,j}}{\sigma_j}\Big) = \sum_{j=1}^{m} \sigma_j^2 \sum_{i=1}^{n} \rho\Big(\frac{t_{i,j} - a_i b_j}{\sigma_j}\Big) \tag{3.4}$$

where $\sigma_j (1 \leq j \leq m)$ is a column-scale parameter and $\rho(x)$ is a bounded ρ-function.

The M-estimators of $\mathbf{a}$, $\mathbf{b}$ and σ can be found by differentiating the function $l(\mathbf{a}, \mathbf{b}, \sigma)$ in equation 3.4, with respect to the each parameter, setting the partial derivatives equal to zero and solving. So the estimators are the solution to the following system of equations:

$$\frac{\partial l(\mathbf{a}, \mathbf{b}, \sigma)}{\partial a_i} = \sum_{j=1}^{m} \sigma_j \rho'\Big(\frac{r_{i,j}}{\sigma_j}\Big) b_j = 0, \qquad i = 1, ..., n; \tag{3.5}$$

$$\frac{\partial l(\mathbf{a}, \mathbf{b}, \sigma)}{\partial b_j} = \sum_{i=1}^{n} \sigma_j \rho'\Big(\frac{r_{i,j}}{\sigma_j}\Big) a_i = 0, \qquad j = 1, ..., m; \tag{3.6}$$

$$\frac{\partial l(\mathbf{a}, \mathbf{b}, \sigma)}{\partial \sigma_j} = 2\sigma_j \sum_{i=1}^{n} \rho\Big(\frac{r_{i,j}}{\sigma_j}\Big) - \sum_{i=1}^{n} r_{i,j} \rho'\Big(\frac{r_{i,j}}{\sigma_j}\Big) = 0, \qquad j = 1, ..., m \tag{3.7}$$

where $\rho'(x) = \frac{d}{dx}\rho(x)$

Let $\psi(x) = \tau\rho'(x)$ with τ a constant. Following the idea stated in Section 2.2.3 of Maronna et al. (2006) that a location M-estimate can be seen as a weighted mean, we used their definition of $W(t)$ to obtain the weights in the system of equations by letting $w_{i,j} = W(r_{i,j}/\sigma_j)$. $W(t)$ is defined as:

$$W(t) = \begin{cases} \frac{\psi(t)}{t} & \text{if } t \neq 0, \\ \psi'(0) & \text{if } t = 0. \end{cases} \tag{3.8}$$

Then the weights will be:

$$w_{i,j} = \begin{cases} \frac{\sigma_j}{r_{i,j}} \psi\Big(\frac{r_{i,j}}{\sigma_j}\Big) & \text{if } \frac{r_{i,j}}{\sigma_j} \neq 0, \\ \psi'(0) & \text{if } \frac{r_{i,j}}{\sigma_j} = 0. \end{cases} \tag{3.9}$$

Maronna and Yohai (2008) proposed the use of an estimate for σ_j ($\hat{\sigma_j}$) instead of solving equation (3.7). They used robust estimates of column variability for $\hat{\sigma_j}$ and fixed these values through out the iterative process. This assumption reduces the

system of equations to only two sets of equations. Then substituting equation (3.9) in equations (3.5) and (3.6) we obtain:

$$\sum_{j=1}^{m} w_{i,j} r_{i,j} b_j = 0 \Longrightarrow \sum_{j=1}^{m} w_{i,j}(t_{i,j} - a_i b_j) b_j = 0 \Longrightarrow \hat{a}_i = \frac{\sum_{j=1}^{m} w_{i,j} t_{i,j} b_j}{\sum_{j=1}^{m} w_{i,j} b_j^2} \tag{3.10}$$

which is the parameter estimator of a weighted simple linear regression through the origin model $t_{i,j} = a_i w_{i,j} b_j$ for $j = 1, ..., m$ and

$$\sum_{i=1}^{n} w_{i,j} r_{i,j} a_i = 0 \Longrightarrow \sum_{i=1}^{n} w_{i,j}(t_{i,j} - a_i b_j) a_i = 0 \Longrightarrow \hat{b}_j = \frac{\sum_{i=1}^{n} w_{i,j} t_{i,j} a_i}{\sum_{i=1}^{n} w_{i,j} a_i^2} \tag{3.11}$$

which is the parameter estimation of a weighted simple linear regression through the origin model $t_{i,j} = b_j w_{i,j} a_i$ for $i = 1, ..., n$.

Also from (3.9) we have that the weights are calculated using the following expression:

$$w_{i,j} = \frac{\hat{\sigma}_j}{r_{i,j}} \tau \rho' \left(\frac{r_{i,j}}{\hat{\sigma}_j} \right) \tag{3.12}$$

The alternating algorithm proposed by Maronna and Yohai (2008) uses equations (3.10), (3.11) and (3.12) to estimate the vectors $\mathbf{a}$ and $\mathbf{b}$. Their iterative algorithm starts with the selection of initial values $(\mathbf{a^0}, \mathbf{b^0})$. With these values $\hat{\sigma_j}$ is estimated for $j = 1, ..., m$. Then the residuals $r_{i,j}$ and the weights $w_{i,j}^0$ are estimated for $i = 1, ..., n$ and $j = 1, ..., m$. With these weights and $\mathbf{b^0}$ a new estimate $\mathbf{a^1}$ is obtained by substituting in equation (3.10). Then a new estimate $\mathbf{b^1}$ is obtained by substituting in equation (3.11) $\mathbf{a^1}$ and $w_{i,j}^0$ for $i = 1, ..., n$ and $j = 1, ..., m$. Then new residuals are computed as well as new weights $w_{i,j}^1$ for $i = 1, ..., n$ and $j = 1, ..., m$. The iterative process continues until either the relative decrease in the loss function is less than a prescribed tolerance value or the number of iterations exceeds a given limit.

3.1.1 Loss function decreasing at each iteration

The algorithm described at the end of the last subsection converges if the loss function decreases at each iteration. So we have to prove that $l(\mathbf{a^{k+1}}, \mathbf{b^{k+1}}, \sigma^{\mathbf{k+1}}) \leq l(\mathbf{a^k}, \mathbf{b^k}, \sigma^{\mathbf{k}})$.

Proof. If σ_j is estimated with the initial values and fixed through the iterations as suggested in the algorithm proposed by Maronna and Yohai (2008) we have that proving $l(\mathbf{a^{k+1}}, \mathbf{b^{k+1}}, \sigma^{\mathbf{k+1}}) \leq l(\mathbf{a^k}, \mathbf{b^k}, \sigma^{\mathbf{k}})$ becomes $l(\mathbf{a^{k+1}}, \mathbf{b^{k+1}}) \leq l(\mathbf{a^k}, \mathbf{b^k})$. Now as the iterative process does the regression for each parameter independently it is enough to show that:

$$l(\mathbf{a^{k+1}}, \mathbf{b^{k+1}}) \leq l(\mathbf{a^{k+1}}, \mathbf{b^k}) \leq l(\mathbf{a^k}, \mathbf{b^k}) \tag{3.13}$$

It is assumed that $\rho(r)$ is a ρ-function, that the function $W(t)$ defined in equation (3.8) is a non-increasing function of $|t|$, and that ψ is continuous.

We will now concentrate on the second inequality. Using equation (3.4) we obtain:

$$\begin{aligned} & l(\mathbf{a^{k+1}}, \mathbf{b^k}) - l(\mathbf{a^k}, \mathbf{b^k}) \\ &= \sum_{j=1}^{m} \hat{\sigma_j}^2 \Big[\sum_{i=1}^{n} \rho\Big(\frac{r_{i,j}(\mathbf{a^{k+1}}, \mathbf{b^k})}{\hat{\sigma_j}} \Big) - \sum_{i=1}^{n} \rho\Big(\frac{r_{i,j}(\mathbf{a^k}, \mathbf{b^k})}{\hat{\sigma_j}} \Big) \Big] \\ &= \sum_{j=1}^{m} \hat{\sigma_j}^2 \Big[\sum_{i=1}^{n} g\Big(\Big(\frac{r_{i,j}(\mathbf{a^{k+1}}, \mathbf{b^k})}{\hat{\sigma_j}} \Big)^2 \Big) - \sum_{i=1}^{n} g\Big(\Big(\frac{r_{i,j}(\mathbf{a^k}, \mathbf{b^k})}{\hat{\sigma_j}} \Big)^2 \Big) \Big] \end{aligned} \tag{3.14}$$

when $\rho(r) = g(r^2)$.

Then $g'(r^2) = \frac{1}{2r} W(r)$. Hence $W(r)$ is a non increasing function of $|r|$ if and only if $g'(r)$ is non increasing. So, as $g'(r)$ is non increasing we can say that $g(y) - g(x) \leq$

$g'(x)(y-x)$ for any x and y value. Using this equation we have:

$$
\begin{aligned}
& l(\mathbf{a^{k+1}}, \mathbf{b^k}) - l(\mathbf{a^k}, \mathbf{b^k}) \\
& \leq \sum_{j=1}^{m} \hat{\sigma_j}^2 \Big[\sum_{i=1}^{n} g'\Big(\Big(\frac{r_{i,j}(\mathbf{a^k}, \mathbf{b^k})}{\hat{\sigma_j}} \Big)^2 \Big) \Big(\Big(\frac{r_{i,j}(\mathbf{a^{k+1}}, \mathbf{b^k})}{\hat{\sigma_j}} \Big)^2 - \Big(\frac{r_{i,j}(\mathbf{a^k}, \mathbf{b^k})}{\hat{\sigma_j}} \Big)^2 \Big) \Big] \\
& = \sum_{j=1}^{m} \sum_{i=1}^{n} g'\Big(\Big(\frac{r_{i,j}(\mathbf{a^k}, \mathbf{b^k})}{\hat{\sigma_j}} \Big)^2 \Big) \Big(\Big(r_{i,j}(\mathbf{a^{k+1}}, \mathbf{b^k}) \Big)^2 - \Big(r_{i,j}(\mathbf{a^k}, \mathbf{b^k}) \Big)^2 \Big) \\
& = \frac{1}{2\tau} \sum_{j=1}^{m} \sum_{i=1}^{n} W\Big(\frac{r_{i,j}(\mathbf{a^k}, \mathbf{b^k})}{\hat{\sigma_j}} \Big) \Big(\Big(r_{i,j}(\mathbf{a^{k+1}}, \mathbf{b^k}) \Big)^2 - \Big(r_{i,j}(\mathbf{a^k}, \mathbf{b^k}) \Big)^2 \Big) \\
& = \frac{1}{2\tau} \sum_{j=1}^{m} \sum_{i=1}^{n} w_{i,j} \Big(r_{i,j}(\mathbf{a^{k+1}}, \mathbf{b^k}) - r_{i,j}(\mathbf{a^k}, \mathbf{b^k}) \Big) \Big(r_{i,j}(\mathbf{a^{k+1}}, \mathbf{b^k}) + r_{i,j}(\mathbf{a^k}, \mathbf{b^k}) \Big)
\end{aligned}
\tag{3.15}
$$

Remembering that $r_{i,j} = t_{i,j} - a_i b_j$ we have:

$$
\begin{aligned}
r_{i,j}(\mathbf{a^{k+1}}, \mathbf{b^k}) - r_{i,j}(\mathbf{a^k}, \mathbf{b^k}) &= b_j^k (a_i^k - a_i^{k+1}) \\
r_{i,j}(\mathbf{a^{k+1}}, \mathbf{b^k}) + r_{i,j}(\mathbf{a^k}, \mathbf{b^k}) &= 2t_{i,j} - b_j^k (a_i^k + a_i^{k+1})
\end{aligned}
\tag{3.16}
$$

Substituting in the equation (3.15) we obtain:

$$
\begin{aligned}
& = \frac{1}{2\tau} \sum_{j=1}^{m} \sum_{i=1}^{n} w_{i,j} b_j^k (a_i^k - a_i^{k+1}) \Big(2t_{i,j} - b_j^k (a_i^k + a_i^{k+1}) \Big) \\
& = \frac{1}{2\tau} \sum_{i=1}^{n} (a_i^k - a_i^{k+1}) \Big[\sum_{j=1}^{m} 2w_{i,j} t_{i,j} b_j^k - \sum_{j=1}^{m} w_{i,j} b_j^{2k} (a_i^k + a_i^{k+1}) \Big]
\end{aligned}
\tag{3.17}
$$

From equation (3.10) we have that:

$$
\sum_{j=1}^{m} w_{i,j} r_{i,j} b_j^k = 0 \Longrightarrow \sum_{j=1}^{m} w_{i,j} (t_{i,j} - a_i^{k+1} b_j^k) b_j^k = 0 \Longrightarrow \sum_{j=1}^{m} w_{i,j} t_{i,j} b_j^k = \sum_{j=1}^{m} w_{i,j} a_i^{k+1} b_j^{2k}
\tag{3.18}
$$

Substituting in the equation (3.17)

$$\begin{aligned}
&= \frac{1}{2\tau}\sum_{i=1}^{n}(a_i^k - a_i^{k+1})\Big[\sum_{j=1}^{m} 2w_{i,j}a_i^{k+1}b_j^{2k} - \sum_{j=1}^{m} w_{i,j}b_j^{2k}(a_i^k + a_i^{k+1})\Big] \\
&= \frac{1}{2\tau}\sum_{i=1}^{n}(a_i^k - a_i^{k+1})\Big[\sum_{j=1}^{m} w_{i,j}b_j^{2k}(a_i^{k+1} - a_i^k)\Big] \\
&= \frac{1}{2\tau}\sum_{j=1}^{m}\sum_{i=1}^{n} w_{i,j}b_j^{2k}(a_i^k - a_i^{k+1})(a_i^{k+1} - a_i^k)
\end{aligned} \tag{3.19}$$

So we have

$$\begin{aligned}
&l(\mathbf{a^{k+1}}, \mathbf{b^k}) - l(\mathbf{a^k}, \mathbf{b^k}) \leq \frac{1}{2\tau}\sum_{j=1}^{m}\sum_{i=1}^{n} w_{i,j}b_j^{2k}(a_i^k - a_i^{k+1})(a_i^{k+1} - a_i^k) \leq 0 \\
&l(\mathbf{a^{k+1}}, \mathbf{b^k}) \leq l(\mathbf{a^k}, \mathbf{b^k})
\end{aligned} \tag{3.20}$$

A similar procedure is followed to prove $l(\mathbf{a^{k+1}}, \mathbf{b^{k+1}}) \leq l(\mathbf{a^{k+1}}, \mathbf{b^k})$. □

This proves that the loss function $l(\mathbf{a}, \mathbf{b})$ decreases at each iteration, and the alternating algorithm proposed by Maronna and Yohai (2008) will converge to a local minimum.

3.2 Asymmetric objective function

The selection of the correct objective function to make a statistical estimation process robust depends on the data configuration, the statistical method and the use that will be made of the estimates. The Huber and the Tukey's bisquare (presented in Table 3.1) are two of the more commonly used objective functions in the robust statistics literature, and their performance will depend on the procedure used. For example, Colin et al. (2008) mentioned that their algorithm experienced convergence problems when ρ-functions which are not convex (also called redescending), like Tukey's bisquare, were used in an alternating regression process. For that reason they suggest the use of convex (called monotone) objective functions like the Huber,

logistic or Fair, because they always converge to a local minimum in the alternating regression. In contrast, the algorithm proposed by Maronna and Yohai (2008) for the low-rank approximation of data matrices with elementwise contamination, is defined for any ρ-function that complies with the definition mentioned in Section 3.1, and with $W(t)$ a non-increasing function of $|t|$. This means that the convergence conditions of the algorithm depends on $W(t)$ being non-increasing function of $|t|$, the convexity of the ρ-function not being important. So both the Huber and the bisquare comply with the convergence conditions of the algorithm proposed by Maronna and Yohai (2008).

The bisquare function has the property of assigning weights equal to zero to large outliers, so this means these outliers are not considered in the estimation. This is a property we are interested in, because we would like our estimates to be not influenced at all by the large outliers in the data. After working with this function we noticed that the procedure was eliminating whole rows or columns from the estimations because the corresponding weights were zero. To solve this our first approach was to introduce a constraint on the way the weights were assigned, but this was indirectly modifying the objective function. The second approach was to propose a different objective function that includes in the weight function the constraint of not leaving a row or a column with all their weights equal to zero.

The proposed function is a combination of two functions.

- For the negative residuals we use the least-squares function multiplied by a constant. This definition will assure that large negative residuals are not down weighted. We do not want to down weight negative residuals because under the assumption of asymmetric outliers, negative residuals correspond to non-outlier observations that we want to preserve.

- For the positive residuals the bisquare function is used. As the positive residuals are weighted using the bisquare function we maintain the property that large positive outliers will not be used in the estimation, as their weights will be zero or close to zero.

As the negative residuals will be given weight equal to one, and we expect to have on average half of the observations by row and by column with negative residuals. Then the problem of having rows and/or columns with all their weights equal to zero is eliminated. For the reason given above we think this proposed objective function should perform better than the bisquare when the outliers are asymmetric. It will be shown with simulations in Chapter 4 that this hypothesis appears to be true.

Method	Objective Function	Weight Function
Least-Squares	$\rho_{LS}(r) = r^2$	$w_{LS}(r) = 1$
Huber	$\rho_H(r) = \begin{cases} \frac{1}{2}r^2 & \lvert r\rvert \leqslant c \\ c\lvert r\rvert - \frac{1}{2}c^2 & \lvert r\rvert > c \end{cases}$	$w_H(r) = \begin{cases} 1 & \lvert r\rvert \leqslant c \\ \frac{c}{\lvert r\rvert} & \lvert r\rvert > c \end{cases}$
Bisquare	$\rho_B(r) = \begin{cases} \frac{c^2}{6}\left[1 - \left[1 - \left(\frac{r}{c}\right)^2\right]^3\right] & \lvert r\rvert \leqslant c \\ \frac{c^2}{6} & \lvert r\rvert > c \end{cases}$	$w_B(r) = \begin{cases} \left[1 - \left(\frac{r}{c}\right)^2\right]^2 & \lvert r\rvert \leqslant c \\ 0 & \lvert r\rvert > c \end{cases}$
Asymmetric	$\rho_{ASY}(r) = \begin{cases} 3\left(\frac{r}{c}\right)^2 & r \leqslant 0 \\ \left[1 - \left[1 - \left(\frac{r}{c}\right)^2\right]^3\right] & 0 < r \leqslant c \\ 1 & r > c \end{cases}$	$w_{ASY}(r) = \begin{cases} 1 & r \leqslant 0 \\ \left[1 - \left(\frac{r}{c}\right)^2\right]^2 & 0 < r \leqslant c \\ 0 & r > c \end{cases}$

Table 3.1: Objective functions

Table 3.1 presents the objective and weights functions for four different function approaches. The least squares which is the usual non-robust estimation. The Huber and bisquare functions, which as mentioned before, are two of the most used functions in robust statistics. The fourth function in the table is our proposal of objective function when asymmetric outliers are present.

The determination of the parameter c in each of the functions directly affects the value of the weights, so after choosing the appropriate ρ-function, a careful selection of the parameter c has to be done. The literature suggests setting $c = 1.345$ for the Huber function, and $c = 4.685$ for the bisquare function. These values correspond to 95% efficiency when the real distribution of the observations is the standard normal distribution. In the following subsection we calculate the value of c for the asymmetric objective function, which corresponds to the same 95% level of efficiency.

Figure 3.1 shows how the different objective functions deal with the residuals. The assumption of symmetric residuals for the first three functions can be appreciated. In contrast the proposed asymmetric function behaves similar to the least squares in the left side and similar to the bisquare in the right hand side.

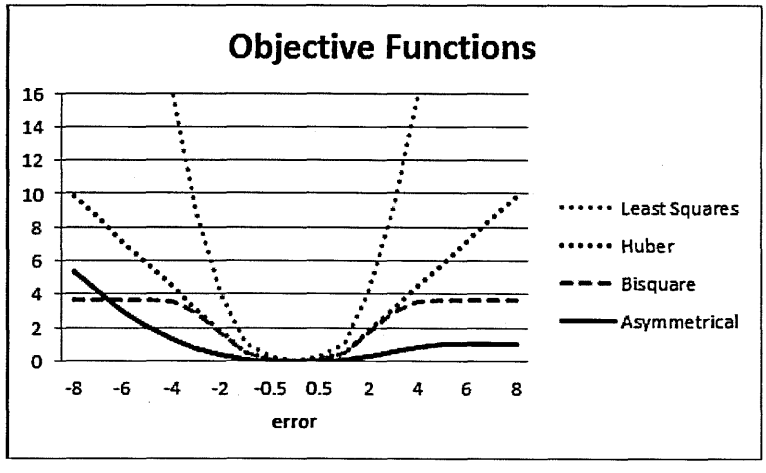


Figure 3.1: Behaviour of the residuals depending on the objective function

The weights associated with each of this functions are specified in Table 3.1. For the least squares objective function the weights are equal to one meaning that all the observations will be considered in the same way. The Huber and bisquare concentrates the estimation on the observations with residuals nearer to zero, giving less weight to those observations with large residuals. The difference between the Huber and bisquare functions is in the way in which the weights decrease as resid-

uals are further from zero. On one hand the Huber gives weights that tend to zero to large residuals, then the weights increases as $|residual|$ gets closer to the interval $(-c, c)$ point at which the weights are one and remain one for all the residuals between $-c$ and c. On the other hand the bisquare gives weights equal to zero to large residuals and remain zero for any residuals outside the interval $(-c, c)$ and then the weights increases as $|residual|$ gets closer to zero, reaching one at $|residual| = 0$.

The asymmetric function gives weights equal to one for negative residuals. This is because we assume that all the outliers are on the right hand side of the distribution so if the residual is negative we are certain it is not an outlier. The function will assign weights zero to positive residuals greater than c, and the weights increases as positive residuals gets closer to zero.

3.2.1 Determination of c

The tuning constant c is chosen in order to ensure a given asymptotic efficiency at the normal distribution. We determine the c value for the asymmetric function based on the procedure given for the Huber and bisquare function in Section 2.2 of Maronna et al. (2006). The asymptotic efficiency of an M-estimate $\hat{\mu}$ of a parameter μ is defined as:

$$\text{Eff}(\hat{\mu}) = \frac{\nu_0}{\nu} \tag{3.21}$$

where ν_0 is the asymptotic variance of the MLE of μ and the asymptotic variance of $\hat{\mu}$ is

$$\nu = \frac{E_{F_0}(\psi(x)^2)}{(E_{F_0}\psi'(x))^2} \tag{3.22}$$

where $\psi(x) = \tau\rho'(x)$ and τ is a constant.

Then from Table 3.1 we obtain that for $\tau = \frac{c^2}{6}$, $\psi_{ASY}(x)$ is:

$$\psi_{ASY}(x) = \begin{cases} x & x \leqslant 0 \\ x\left[1 - \left(\frac{x}{c}\right)^2\right]^2 & 0 < x \leqslant c \\ 0 & x > c \end{cases} \tag{3.23}$$

To calculate the asymptotic efficiency of the asymmetric function at the normal distribution, we set $F_0 = N(0,1)$ and find ν by substituting 3.23 in 3.22.

$$\nu = \frac{\int_{-\infty}^{0} x^2\phi(x)dx + \int_0^c x^2[1-(\frac{x}{c})^2]^4\phi(x)dx}{\left[\int_{-\infty}^{0}\phi(x)dx + \int_0^c [1-(\frac{x}{c})^2]^2 - 4\frac{x^2}{c^2}[1-(\frac{x}{c})^2]\phi(x)dx\right]^2} \tag{3.24}$$

where $\phi(x)$ is the standard normal density.

We estimate the integrals in R for given values of c, and with the results we calculate the correspondent asymptotic variance, hence the asymptotic efficiency for different values of c. The results are presented in the following table.

eff	0.80	0.85	0.90	0.95
c	2.6	3	3.5	4.3

Table 3.2: Values of c for different efficiencies of the asymmetric estimate.

The selection of a value for the tuning constant will depend on the desired trade-off between robustness and efficiency. From Table 3.2 we observe that 95% efficiency when the real distribution is the standard normal distribution is achieved when $c = 4.3$.

3.3 Improving the algorithm

In this section we will describe further modifications made to the algorithm proposed by Maronna and Yohai (2008).

1. Due to the nature of the outliers in the data, small times are less likely to have positive outliers. So we propose initial values that are related to smaller values. (It will be shown in Chapter 4 that for the improved algorithm the choice of initial values is not critical)

2. We also introduce an updating procedure for the scale parameter, so instead of being an estimate made with the initial values, it is estimated in each iteration.

3. The stopping rule used only depends on the number of iterations.

Because the approximation of a matrix by the multiplication of two vectors has an infinite number of solutions, we introduce at every iteration a normalization of the vector $\mathbf{b}$ so that is restricted to have values that add up to a known value D. On the orienteering data context the vector estimates based on this normalization lead to a useful interpretation of the vectors, which is speeds and distances.

Subsection 3.3.4 describes the procedure proposed by Maronna and Yohai (2008) to deal with missing values in the data. We use the same procedure with the specification that there should not be more than half missing values in a row or column. If the matrix has rows or columns with more than half missing values they are dropped from the analysis. Finally the last subsection presents step by step the improved algorithm proposed in this work.

3.3.1 Initial values

Maronna and Yohai (2008) mention the need for robust initial values estimates to avoid falling into "bad" local minima caused by the non convexity of the proposed loss function. They proposed a procedure based on alternate regression using the median-of-slopes estimate to find initial estimates for rank one fits.

Gabriel and Zamir (1979) in their iterative weighted least squares algorithm, suggest the use of the column with the longest weighted norm as initial value. However there is no restriction regarding the use of row and columns interchangeably as initial values. So the row with the longest weighted norm could have been chosen as initial value. The algorithm of Gabriel and Zamir (1979) assumes that the weights are known, and they mention that for the particular case of using these initial values all the weights of the matrix elements should not be zero nor very small.

Our procedure for the selection of initial values should consider: 1) the data have outliers, 2) the constraint on the columns, and 3) the weights are unknown. We propose to use the shortest norm instead of the longest, because the row with the shortest norm will be the least likely to be affected by the outliers. It is also considered that the row selected should not have missing values. As the weights are unknown we will use the median by row and column to estimate the vectors $\mathbf{a}$ and $\mathbf{b}$ initial values. This procedure is similar to the one used by Maronna and Yohai (2008) to obtain initial values.

To find the initial values $\mathbf{a}^0$ and $\mathbf{b}^0$, first we set D as the value to which all the elements of the vector $\mathbf{b}$ have to add up. Then we distribute D according to the values of the row with the shortest norm, and we name this $\mathbf{b}^{00}$. In other words

$$\mathbf{b}^{00} = (t_{1,1}, ..., t_{1,m}) \times \frac{D}{\sum_{j=1}^{m} t_{1,j}} \tag{3.25}$$

where $i = 1$ is the row with no missing values for which $\sum_{j=1}^{m} t_{i,j}$ is minimised. Then our initial estimate of the vector $\mathbf{a}$, $\mathbf{a}^0$ will be calculated using median values in the following way. The ith element of the vector is estimated by: $a_i^0 = med(\frac{t_{i,1}}{b_1^{00}}, \frac{t_{i,2}}{b_2^{00}}, ..., \frac{t_{i,m}}{b_m^{00}})$. With the vector $\mathbf{a}^0$ we repeat the procedure but now for the

other vector. The vector $\mathbf{b}^0$ is estimated in two steps, first we estimate the jth element of the vector with b_j by $med(\frac{t_{1,j}}{a_1^0}, \frac{t_{2,j}}{a_2^0}, ..., \frac{t_{n,j}}{a_n^0})$, then the vector $\mathbf{b}^0$ is calculated by normalizing $\mathbf{b}$ so that the values on the vector add up to D.

In the next chapter we will compare the estimates obtained with this initial value procedure against the estimates obtained with random starts. We will show that the use of our initial values appears to produce final estimates that are very similar to the ones obtained using a random start. So the choice of initial value is not critical. The simulation study in the next chapter also shows that both estimates are close to the real values.

3.3.2 Scale parameter $\hat{\sigma_j}(1 \leq j \leq m)$

The estimation of $\hat{\sigma_j}$ for $j = 1, ..., m$ is done with the residuals ($r_{i,j} = t_{i,j} - \hat{a_i}\hat{b_j}$, $i = 1, ..., n$ $j = 1, ..., m$). This estimate, like the initial values for $\mathbf{a}$ and $\mathbf{b}$, has to be robust.

Based on the simultaneous M-estimates mentioned in Section 2.6 of Maronna et al. (2006), the minimization of equation (3.4) uses equations (3.5) and (3.6) to estimate the vectors $\mathbf{a}$ and $\mathbf{b}$ respectively. For the scale parameter the system of equations used is not the one defined in (3.7), instead Maronna et al. (2006) used the following approach:

$$\frac{1}{n}\sum_{i=1}^{n}\rho_{scale}\left(\frac{r_{i,j}}{\hat{\sigma_j}}\right) = \delta. \tag{3.26}$$

They also mentioned that a very robust estimator choice of ρ_{scale} is the step function $\rho_{scale}(t) = I(|t| > c)$ with $\delta = .5$ and $c = 0.675$ to make it consistent for the standard deviation when the underlying distribution it the standard normal distribution. So the estimate $\hat{\sigma_j}$ proposed for regression M-estimates in Subsection 4.5.2 Maronna

et al. (2006) is $\hat{\sigma}_j = \frac{1}{0.675} med_i[|r_{i,j}|, r_{i,j} \neq 0]$ which is a regression and affine invariant and scale equivariant estimate.

Maronna and Yohai (2008) used the following equation to compute the scale estimate using the initial residuals.

$$\frac{1}{n}\sum_{i=1}^{n}\rho\Big(\frac{r_{i,j}}{\hat{\sigma}_j}\Big) = \delta. \tag{3.27}$$

with ρ being the same objective function used for the algorithm, that could be one of the presented in Table 3.1, and

$$\delta = \frac{nm - n + 2m}{2nm} \tag{3.28}$$

Then the estimate is normalized to be consistent at the normal distribution and left fixed for the rest of the iterations. As mentioned in Section 3.1 this ensures that the loss function $l(\mathbf{a}, \mathbf{b}, \sigma)$ decreases at each iteration, provided that $W(t)$ is a non-increasing function of $|t|$. Then the alternating algorithm proposed will converge to a local minimum.

Our proposal is to use the estimator Sn, defined by Rousseeuw and Croux (1993) and implemented in R. As mentioned in Subsection 2.3.1, this scale estimator does not use a location estimate and for that reason according to Rousseeuw and Croux (1993) the Sn is more efficient than MAD under asymmetric distributions. The estimator Sn was defined in Subsection 2.3.1. So the scale estimate $\hat{\sigma}_j$ in the loss function (3.4) is:

$$\hat{\sigma}_j = Sn_j = 1.1926 \times med_i\{med_{l:l\neq i}|r_{i,j} - r_{l,j}|\} \tag{3.29}$$

where $r_{i,j}$ are the residuals.

Maronna et al. (2006) in Section 4.5.2 mention that the scale parameter $\hat{\sigma}$ has to be updated at each iteration when using an iterative reweighting method to simultaneously estimate the parameters of the regression and the scale parameter.

Analysing the convergence of the algorithm, we have found out that if the scale parameter is estimated by Sn_j but not updated at each step we can use the same argument as Maronna and Yohai (2008), or the results of alternating minimization for non-negative matrix approximation, to prove that our algorithm converges to a local minimum. However if the scale parameter is updated at each iteration, the loss function $l(\mathbf{a}, \mathbf{b}, \sigma)$ does not necessarily decrease at each iteration, so it might not reach the minimum value. However, we will show with simulations in Chapter 4 that in practice and in the cases we studied the point reached by the loss function with our algorithm appears to be at least as good as the point reached with the method proposed by Maronna and Yohai (2008).

3.3.3 Stopping rule

The iterative procedure proposed by Maronna and Yohai (2008) stops when either the relative decrease in the loss function (3.4) is less than 0.001, which is the specified tolerance value, or the number of iterations exceeds a given limit of 20.

Our algorithm will stop only after 20 iterations have been performed. This rule appears to be effective to find good estimates of $\mathbf{a}$ and $\mathbf{b}$. Also it seems the estimate of the scale parameter and the weights reach a stable point when 20 iterations are performed. This means that despite the starting values used, the estimates produced by the algorithm converge to the same values.

In the next chapter the stabilizing effect on the estimates and the weights is shown. This is done for a simulated data set and using a number of random starting values. In all the cases, after 20 iterations the algorithm end at very similar points.

3.3.4 Missing values

Maronna and Yohai (2008) adapted their procedure for the case of scattered missing values, when these missing values are a small proportion in any row or column. We use the same modification in our algorithm for data sets that have information in at least half of the columns per row and half of the rows per column.

The loss function is calculated with:

$$l(\mathbf{a}, \mathbf{b}, \sigma) = \sum_{j=1}^{m} \sigma_j^2 \sum_{i=1}^{n} M_{i,j} \rho\Big(\frac{r_{i,j}}{\sigma_j}\Big) \tag{3.30}$$

where $M_{i,j}$ is the indicator that $t_{i,j}$ is not missing.

This indicator affects the weights function. It will assign weights equal to zero to all the elements in the matrix that are missing values.

3.3.5 Improved M&Y algorithm

This subsection presents step by step the algorithm proposed in this work to reduce a matrix $\mathbf{T}$ with dimensions $n \times m$ into two vectors $\mathbf{a}$ and $\mathbf{b}$ with dimensions n and m respectively. This algorithm is general for any ρ-function.

Set

$\mathbf{t}_{i,\cdot}$ as the i-th row of the matrix $\mathbf{T}$.

$\mathbf{t}_{\cdot,j}$ as the j-th column of the matrix $\mathbf{T}$.

D as the value of the sum which all the elements of the vector $\mathbf{b}$ has to take.

1. Find the row with the shortest norm without missing values $\mathbf{t}_{min,.}$. Calculate the starting value $\mathbf{b}^{00}$ as D distributed according to the row with the shortest norm.

$$\mathbf{b}^{00} = \frac{D}{\sum_{j=1}^{m} t_{min,j}} \times \mathbf{t}_{min,.} \tag{3.31}$$

2. Estimate $\mathbf{a}^0 = (a_1^0, a_2^0, ..., a_n^0)$ using the following equation: $a_i^0 = med(\frac{\mathbf{t}_{i,.}}{\mathbf{b}^{00}})$ for $i = 1, ..., n$

3. (a) Estimate $\mathbf{b} = (b_1, b_2, ..., b_m)$ using the following equation: $b_j = med(\frac{\mathbf{t}_{.,j}}{\mathbf{a}^0})$ for $j = 1, ..., m$

   (b) Set $\mathbf{b}^0 = (b_1^0, b_2^0, ..., b_n^0) = \frac{D}{\sum_{j=1}^{m} b_j} \times \mathbf{b}$

4. Compute the residuals matrix $\mathbf{R}^0 = \mathbf{T} - \mathbf{a}^0\mathbf{b}^{0t}$

5. Estimate $\hat{\sigma}_j^0 = Sn_j = 1.1926 \times med_i\{med_l|r_{i,j} - r_{l,j}|\}$ for $j = 1, ..., m$

6. Calculate the loss function

$$l(\mathbf{a}^0, \mathbf{b}^0, \hat{\sigma}^0) = \sum_{j=1}^{m} \hat{\sigma_j}^2 \sum_{i=1}^{n} \rho\left(\frac{r_{i,j}}{\hat{\sigma}_j}\right) \tag{3.32}$$

   using the selected ρ-function.

7. Using the following equation

$$w_{i,j} = \frac{\hat{\sigma}_j}{r_{i,j}} \tau\rho'\left(\frac{r_{i,j}}{\hat{\sigma}_j}\right) \tag{3.33}$$

   estimate the weights $w_{i,j}^0$ for $i = 1, ..., n$ and $j = 1, ..., m$. If $t_{i,j}$ is a missing value set its weight $w_{i,j}^0$ to be zero.

8. Do iterations. That is, for k=1 to 20

   (a) Substitute the weights $w_{i,j}^{k-1}$ and the vector $\mathbf{b}^{k-1}$ in following equation:

$$\hat{a}_i = \frac{\sum_{j=1}^{m} w_{i,j} t_{i,j} b_j}{\sum_{j=1}^{m} w_{i,j} b_j^2} \tag{3.34}$$

   and hence estimate $\mathbf{a}^k = (a_1^k, a_2^k, ..., a_n^k)$ using weighted linear regression.

   (b) i. Substitute the weights $w_{i,j}^{k-1}$ and the vector $\mathbf{a}^k$ in the following equation:

$$\hat{b}_j = \frac{\sum_{i=1}^{n} w_{i,j} t_{i,j} a_i}{\sum_{i=1}^{n} w_{i,j} a_i^2} \tag{3.35}$$

   and hence estimate $\mathbf{b} = (b_1, b_2, ..., b_m)$, using weighted linear regression.

   ii. Set $\mathbf{b}^k = (b_1^k, b_2^k, ..., b_m^k) = \frac{D}{\sum_{j=1}^{m} b_j} \times \mathbf{b}$

   (c) Compute the residuals matrix $\mathbf{R}^k = \mathbf{T} - \mathbf{a}^k \mathbf{b}^{k^t}$

   (d) Estimate $\hat{\sigma}_j^k = Sn_j = 1.1926 \times med_i\{med_l|r_{i,j} - r_{l,j}|\}$ for $j = 1, ..., m$

   (e) Calculate the value of the loss function $l(\mathbf{a}^k, \mathbf{b}^k, \hat{\sigma}^k)$.

   (f) Using the following equation

$$w_{i,j} = \frac{\hat{\sigma}_j}{r_{i,j}} \tau \rho' \left(\frac{r_{i,j}}{\hat{\sigma}_j} \right) \tag{3.36}$$

   estimate the weights $w_{i,j}^k$ for $i = 1, ..., n$ and $j = 1, ..., m$. If $t_{i,j}$ is a missing value set its weight $w_{i,j}^k$ to be zero.

As mentioned this algorithm performs with any ρ-function, this objective function can be symmetric like the least squares, Huber or bisquare functions. Motivated by the type of data observed in orienteering, we defined an asymmetric ρ-function to be used in our algorithm. In the following chapter we will see the results of using

an asymmetric function instead of a symmetric one, when the data has asymmetric outliers.

This algorithm differs from the algorithm proposed by Maronna and Yohai (2008) in the initial values selection, the scale parameter estimation and the stopping rule. In the following chapter we will show through simulations how these modifications improve the algorithm proposed by Maronna and Yohai (2008).

Chapter 4

Testing the algorithm

This chapter presents the results of analysing the performance of the algorithm that we developed in Chapter 3, that we named Improved M&Y algorithm. The testing is done through simulated data sets. The simulation process is described in Section 4.1. We analyse the performance of the proposed asymmetric objective function compared with the least square, Huber and bisquare functions presented in Chapter 3. The results of that analysis are presented in Section 4.3 and 4.4. Then we move on to study the advantages of the modifications done to the algorithm proposed by Maronna and Yohai (2008). In Section 4.5 we also discuss the convergence of the algorithm to minimal values of the loss function. The last section, Section 4.6, presents an analysis of how sensitive is the algorithm to data sets that do not have clear asymmetric outliers.

4.1 Simulated data

Simulated data sets will be used to test the performance of the Improved M&Y algorithm. The simulated data sets come from the multiplication of two positive vectors called $\mathbf{a}$ and $\mathbf{b}$, then errors $(e_{i,j})$ and positive asymmetric outliers $(u_{i,j})$ are

added. This means that each element of the simulated data set is of the form:

$$t_{i,j} = a_i b_j + e_{i,j} + u_{i,j} I_{i,j}^{\{0,1\}} \tag{4.1}$$

Doing the simulation this way makes possible a comparison between the estimates obtained from the algorithm and the true vector values, and we use this comparison to assess the accuracy of the estimates.

The vectors, the errors and the outliers used in the simulations can be sampled from a wide range of different distributions. Also the parameter values can be chosen from all the options available according to the distribution. The options chosen for the simulations were as follows:

Set

1. $a_i \sim$ Uniform(10,20), where a_i is the i-th element of the vector $\mathbf{a}$.
2. $b_j \sim$ Uniform(0,0.5) where b_j is the j-th element of the vector $\mathbf{b}$.
3. $e_{i,j} \sim N(0, \sigma^2)$ for $i = 1, ..., n$, $j = 1, ..., m$ where σ^2 will be one or b_j^2.
4. $u_{i,j} \sim$ Uniform(5,30). This represents the asymmetric outliers that will be assigned randomly with an indicator function.
5. $I_{i,j}^{\{0,1\}}$ will be 1 with probability $s/100$. This means that the matrix will on average have $s\%$ of outliers.

With this selection we are assured to obtain a positive matrix with asymmetric outliers that comes from the multiplication of two positive vectors. We are interested in these types of data set because it is similar to the orienteering time matrices that we are studying. For this simulation study we have also set the conditions of $n = 50$, $m = 15$ and $s = 20$. These values were selected so that the simulated data sets

resemble a common orienteering data set.

4.1.1 Methodology to assess algorithm's estimates

The analysis is based on N=1000 simulated data sets. The chosen algorithm is applied to obtain estimates of the two vectors that produce the original matrix without outliers. The normalized bias (NBIAS) and the normalized mean square error (NMSE) for the estimates of both vectors are calculated.

The mean normalized bias of the vector **a** estimate ($\hat{\mathbf{a}}$) is calculated as the mean of the following matrix $\mathbf{D_a}$.

$$\mathbf{D_a} = \begin{pmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,N} \\ d_{2,1} & d_{2,2} & \dots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n,1} & d_{n,2} & \dots & d_{n,N} \end{pmatrix}$$

where each element of the matrix is of the form $d_{i,z} = \frac{\hat{a}_i^z - a_i}{a_i}$ for the i-th element of the vector **a** from the z-th simulated data set. Using the square of each element of the matrix $\mathbf{D_a}$ we calculate the NMSE for the vector estimate. A similar procedure is done to obtain the matrix $\mathbf{D_b}$ and calculate the normalized bias and NMSE for the estimate ($\hat{\mathbf{b}}$) of the vector **b**.

4.2 Selection of the minimization rank

The method proposed by Maronna and Yohai (2008) and discussed in Section 3.1 allows the approximation of the matrix **T** by any two matrices with dimensions $\mathbf{A_{(nxp)}}$ and $\mathbf{B_{(pxm)}}$ with $p < min(n, m)$. So when an approximation is made the value of p has to be chosen. As mentioned in Chapter 2 for the case of the orien-

teering times, if the rank of the minimization is $p = 1$ then the results are easier to interpret by relating the matrix $\mathbf{A}$ to the orienteers' speed and the matrix $\mathbf{B}$ to the legs' distance. However we would like to investigate whether the use of a greater rank value (p) produces better fits to the original matrix. This section compares the results of using a lower rank approximation algorithm with rank one ($p = 1$) against the approximation using a rank of two ($p = 2$).

In order to perform the comparison we simulate the matrix $\mathbf{T}$ of dimensions $n \times m$ as the multiplication of:

1. Two rank 1 matrices of the form $\mathbf{A_{(nx1)}}$ and $\mathbf{B_{(1xm)}}$.

2. Two rank 2 matrices of the form $\mathbf{A_{(nx2)}}$ and $\mathbf{B_{(2xm)}}$.

The rank 2 matrices were simulated as follow:
For each element of the matrix $\mathbf{A}$, $A_{i,z} \sim$ Uniform(10,20) for $i = 1, ..., n$ and $z = 1, 2$.
For each element of the matrix $\mathbf{B}$, $B_{z,j} \sim$ Uniform(0,0.5) for $z = 1, 2$ and $j = 1, ..., m$

This procedure makes unlikely that the columns in the matrix are linearly dependent. The outliers and the errors are simulated once for both cases using the methodology described in Section 4.1.

Then we obtain a rank $p = 1$ approximation ($\hat{\mathbf{T}}_\mathbf{1}$) for both cases by applying the Improved M&Y algorithm described in Subsection 3.3.5. For the rank $p = 2$ approximation ($\hat{\mathbf{T}}_\mathbf{2}$), a modification to the Improved M&Y algorithm is implemented. This modification is based on methodology proposed by Maronna and Yohai (2008) for ranks greater than one. This method uses multiple weighted linear regression to estimate the matrices in the iterative process of the algorithm described in Section 3.1. This method has the following process to obtain the initial values; first

run the regression for the first column as a rank one approximation; then use the residuals as the new matrix to estimate the second column again as a rank one approximation. The methodology is similar for higher rank values.

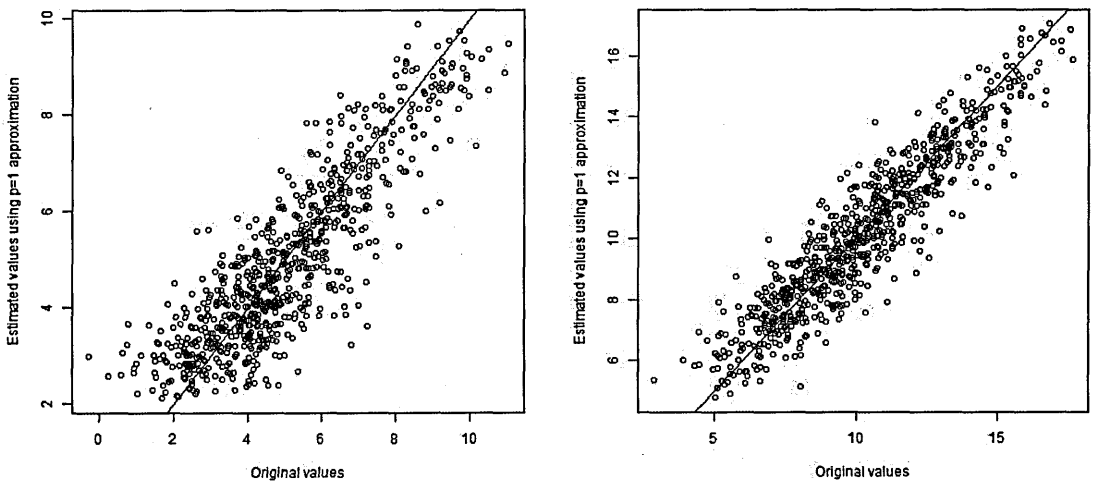
To compare $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$ approximations we simulated 1000 data sets without outliers for each case and computed the residuals. Table 4.1 shows the mean and standard deviation of the sum of the absolute values of the residuals for each case over the 1000 simulations. The table shows that fitting rank two matrices has on average smaller residuals independently of the rank of the original matrices. However for both cases (rank one and rank two original matrices) the means of the absolute residuals for fitting rank one matrices are not considerably larger than the mean values of the rank two fit. Table 4.1 shows similar values of the standard deviation for the sum of the absolute residuals for all the cases, which means that both approximations produced similar results in terms of the average residuals.

ORIGINAL	FITTED			
	Rank=1		Rank=2	
	mean	sd	mean	sd
Rank=1	571.39	16.37	524.23	15.75
Rank=2	588.71	18.13	535.48	15.99

Table 4.1: Mean and standard deviation of the sum of the absolute residuals.

The next step is to compare for one simulated data set the estimates with the original values and assess which of the ranks is producing more accurate estimates.

Figure 4.1 presents the $\hat{\mathbf{T}}_1$ estimates obtained by approximating the matrix $\mathbf{T}$ (simulated from (a) a rank one data set, and (b) a rank two data set) following a rank $p = 1$ procedure. On the plots, each point correspond to one element of the matrix. We observed that the estimates obtained ($\hat{\mathbf{T}}_1$) are close to the original values as the points are near the identity line in both cases, where the identity line corresponds



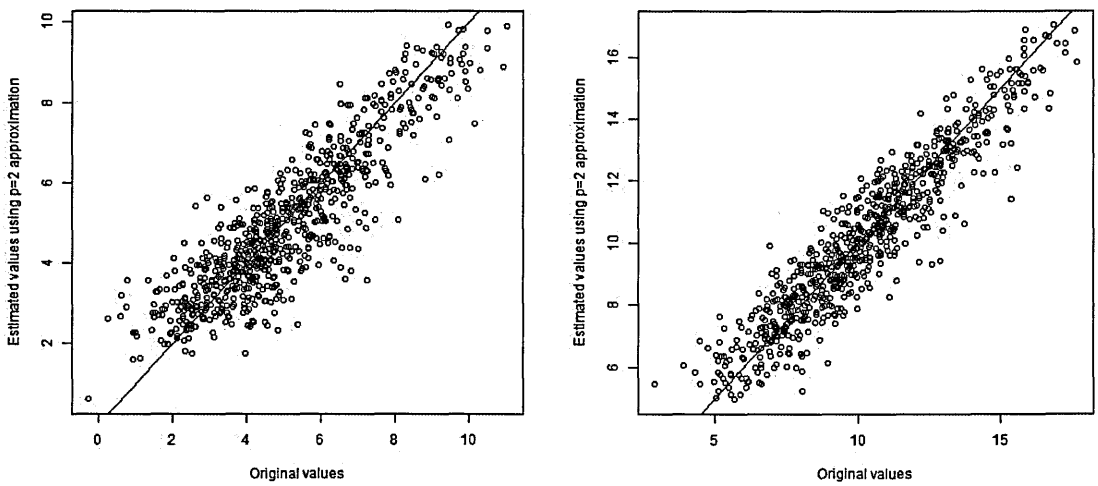
(a) Original simulated from $p = 1$ (b) Original simulated from $p = 2$

Figure 4.1: Comparison of the rank $p = 1$ estimates ($\hat{\mathbf{T}}_1$) against the real values. The line plotted is $y = x$.

to the line $y = x$. This means that for data set coming form a rank one or rank two, the rank one approximation,$\hat{\mathbf{T}}_1$, produces good estimates.

In Figure 4.2 we observe the comparison now of the $\hat{\mathbf{T}}_2$ estimates obtained from the algorithm with rank $p = 2$ against the original values. The estimates obtained ($\hat{\mathbf{T}}_2$) are close to the original values as the points are near the identity line. This means that for data set coming form a rank one or rank two, the rank two approximation,$\hat{\mathbf{T}}_2$, produces good estimates.

Now we compare the estimated matrix of rank one ($\hat{\mathbf{T}}_1$) with the estimated matrix of rank two ($\hat{\mathbf{T}}_2$). If both lower rank approximations perform similarly ($\hat{\mathbf{T}}_1 \approx \hat{\mathbf{T}}_2$) we expect the observations to lie near the identity line in each of the plots.

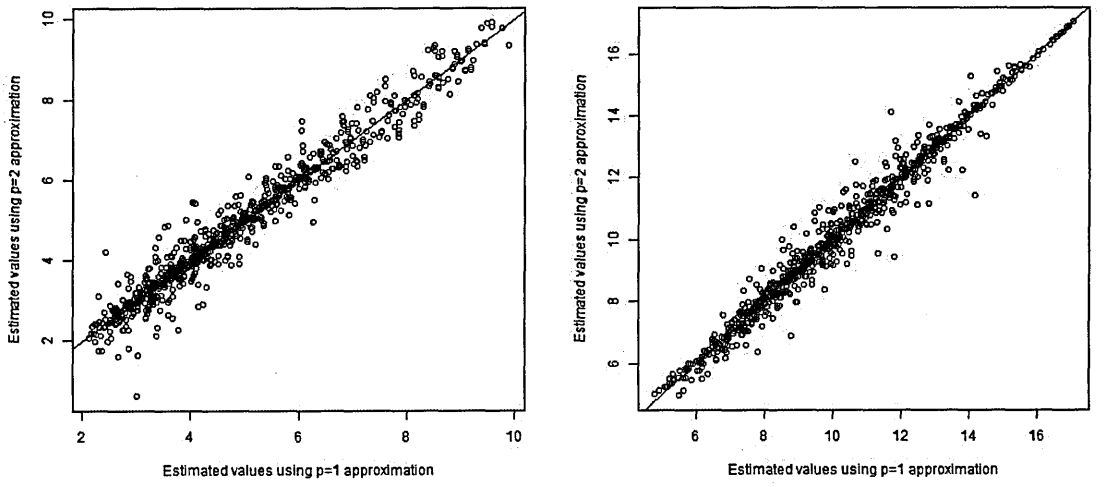


(a) Original simulated from $p = 1$ (b) Original simulated from $p = 2$

Figure 4.2: Comparison of the rank $p = 2$ estimates ($\hat{\mathbf{T}}_2$) against the real values. The line plotted is $y = x$.

Figure 4.3 presents in the first graph the data set simulated as the multiplication of two rank one matrices, the x-axis represents the matrix estimates from the multiplication of two rank one matrices ($\hat{\mathbf{T}}_1$), and the y-axis represents the matrix estimates from the multiplication of two rank two matrices ($\hat{\mathbf{T}}_2$). Each point in the plot represents an element of the original matrix $\mathbf{T}$, which have a correspondent estimate in $\hat{\mathbf{T}}_1$ and another in $\hat{\mathbf{T}}_2$. Similarly the second graph is for data simulated from two rank two matrices. The diagonal lines in the plots represent the identity function.

Figure 4.3 shows that the estimates obtained from a rank one algorithm seem to be similar to the estimates from a rank two algorithm. This is because all the points lie close to the identity line (the diagonal line added in each plot). This indicates that even when the data come from the multiplication of two rank two matrices with no outliers, an approximation with two rank one matrices is similar to an approximation with two rank two matrices.



(a) Original simulated from $p = 1$ (b) Original simulated from $p = 2$

Figure 4.3: Comparison of approximating with $p = 1$ vs $p = 2$ without outliers. The line plotted is $y = x$.

The previous results are based on simulations with no outliers. Now we analyse the impact the outliers have on the estimates. So we repeat the last analysis but with a data set with outliers. Table 4.2 shows the mean and standard deviation of the sum of the absolute values of the residuals for each case over the 1000 simulations.

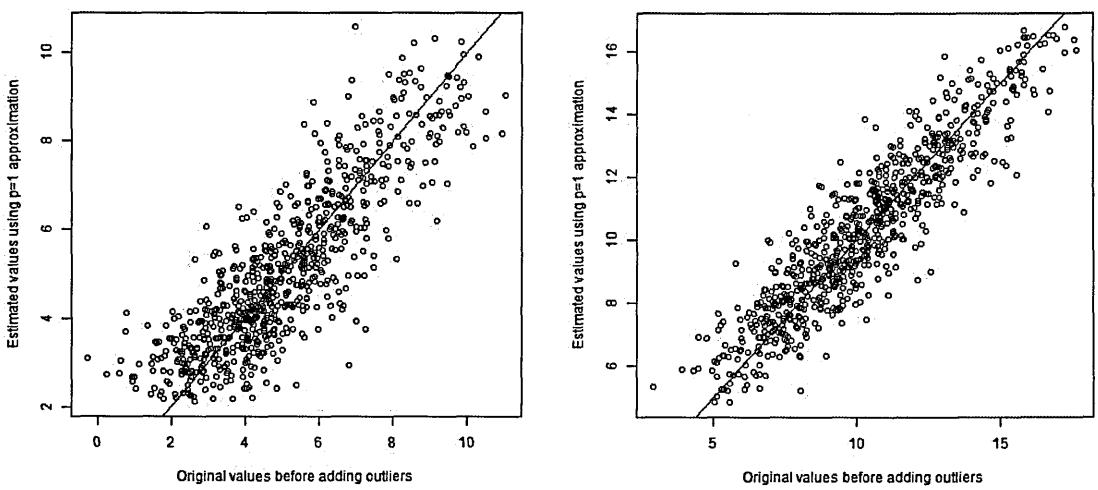
These data sets were simulated with outliers, and it is expected that the objective function in the algorithm will detect those outliers and not consider them in the estimation. For that reason when we compare the estimates against the original values to calculate residuals or produce comparative plots, we use as original values the simulated data set without outliers.

	FITTED			
ORIGINAL	Rank=1		Rank=2	
	mean	sd	mean	sd
Rank=1	593.10	25.36	1028.99	310.47
Rank=2	611.10	25.17	1045.61	347.89

Table 4.2: Mean and standard deviation of the sum of the absolute residuals.

From Table 4.2 we can say that if the data come from the multiplication of two rank one matrices, the residuals obtained by fitting two rank one matrices are on average smaller than the residuals obtained by approximating the original matrix with two rank two matrices. A similar result is observed for the case when the matrix to be reduced comes from the multiplication of two rank two matrices. Also the standard deviations for the rank two approximations are higher than the rank one approximations. This suggests that compared with a rank one approximation, the rank two approximation might be over fitting the model and as a consequence it fails to identify the outliers and hence produces a very bad approximation.

The next step is to compare for one simulated data set the estimates with the original values and assess which of the ranks is producing more accurate estimates.



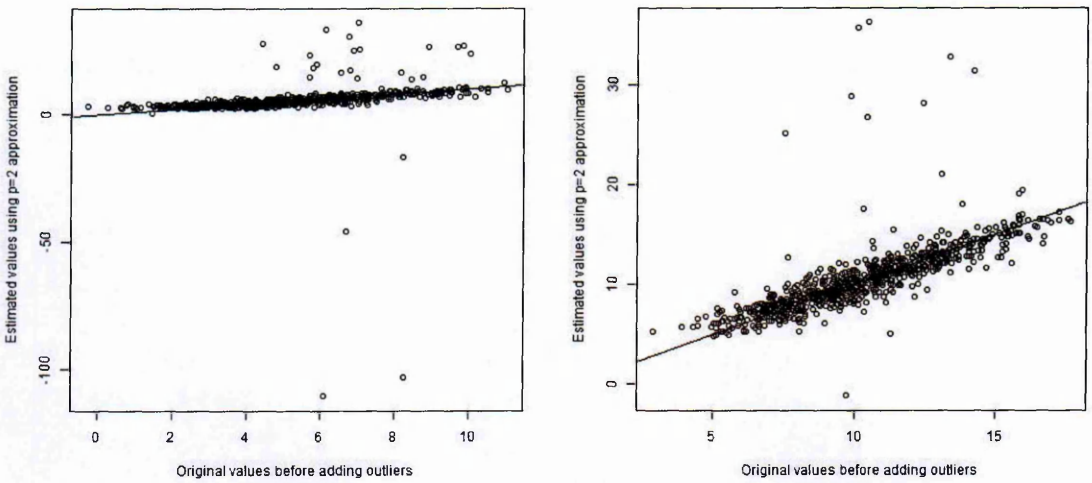
(a) Original simulated from $p = 1$ (b) Original simulated from $p = 2$

Figure 4.4: Comparison of the rank $p = 1$ estimates ($\hat{\mathbf{T}}_1$) against the real values. The line plotted is $y = x$.

Figure 4.4 presents the $\hat{\mathbf{T}}_1$ estimates obtained by approximating the matrix $\mathbf{T}$ (simulated from (a) a rank one data set, and (b) a rank two data set) following a rank $p = 1$ procedure. On the plots, each point corresponds to one element of the matrix,

so there are 750 points, and 150 (20%) of them are outliers. We observe that the estimates obtained ($\hat{\mathbf{T}}_1$) are close to the original values as the points are near to the identity line. This means that for data set of rank one or rank two, the rank one approximation, $\hat{\mathbf{T}}_1$, produces good estimates.

In Figure 4.5 we observe the comparison now with the $\hat{\mathbf{T}}_2$ estimates obtained from the algorithm with rank $p = 2$. We observe that some of the estimated values are significantly higher or smaller than the real values, suggesting that the procedure is not detecting correctly the outliers in the data, affecting the estimates.



(a) Original simulated from $p = 1$ (b) Original simulated from $p = 2$

Figure 4.5: Comparison of the rank $p = 2$ estimates ($\hat{\mathbf{T}}_2$) against the real values. The line plotted is $y = x$.

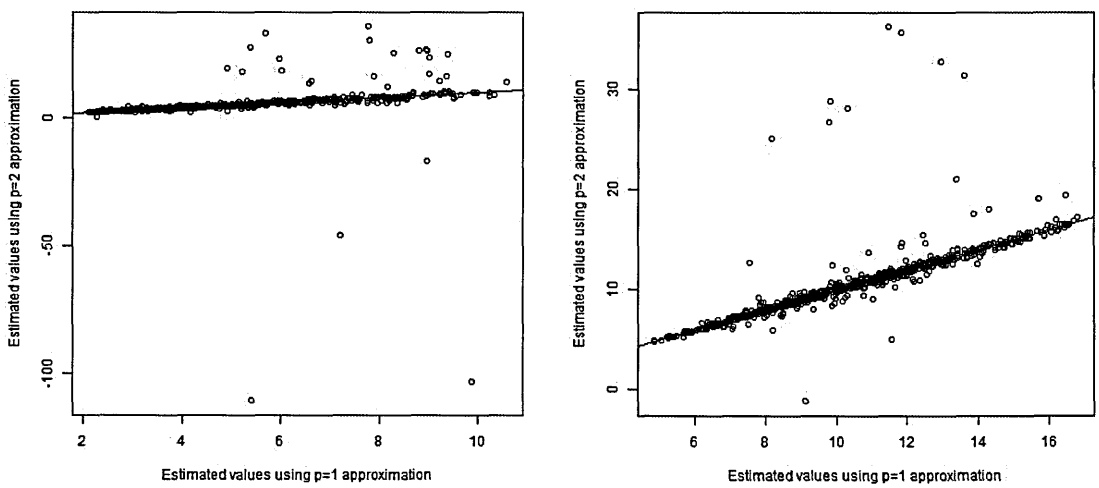
From Figures 4.4 and 4.5 it appears that for matrices with asymmetric outliers, the lower rank approximation with matrices of rank one produces better estimates than the approximation with two matrices of rank two, independently of the rank used to generate the original data set.

Table 4.3 shows for these particular data sets, the sum of the absolute values of the residuals for each case. It can be seen that the values are in the interval formed by the mean and standard deviation for each case presented in Table 4.2.

ORIGINAL	FITTED	
	Rank=1	Rank=2
Rank=1	634.03	1206.67
Rank=2	642.90	805.22

Table 4.3: Sum of the absolute residuals.

Figure 4.6 presents the comparison of $\hat{\mathbf{T}}_1$ against $\hat{\mathbf{T}}_2$ for a data set simulated from (a) two rank one matrices and (b) two rank matrices.



(a) Original simulated from $p = 1$ (b) Original simulated from $p = 2$

Figure 4.6: Comparison of approximating with $p = 1$ vs $p = 2$ with outliers. The line plotted is $y = x$.

The plots show that when the data has outliers there will be a set of points in the estimated matrices $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$ that differ significantly between each other. The majority of the data estimates seem to be fine, however as we are interested in detecting outliers we want an estimation method that is robust to their presence and does not over-fit as seems to be the case for the $\hat{\mathbf{T}}_2$ estimate.

Going back to the purpose of the model, we want to use the estimated speeds to measure the fitness performance and to detect the outliers correctly. Considering, first, that lower rank approximation with $p = 2$ will be harder to interpret, and second, that the estimates do not seem to be better than the estimate from a rank one approximation, even if the underlying model is actually $p = 2$, we conclude that there is no need of a more complicated model, and the use of a lower rank approximation with $p = 1$ is a good choice.

4.3 Performance of the asymmetric objective function

This section presents the comparison between the estimates obtained from the algorithm proposed by Maronna and Yohai (2008) with the asymmetric objective function against the results obtained using least squares, Huber and bisquare symmetric objective functions.

The data used in this analysis was simulated as described in Section 4.1. We are interested in the estimates of the vectors, so we follow the methodology in Subsection 4.1.1 to assess the accuracy of the estimates obtained with the algorithm proposed by Maronna and Yohai (2008).

We use two different distributions for the errors: Case A, $e_{i,j} \sim N(0, 1)$ and Case B, $e_{i,j} \sim N(0, b_j^2)$. We introduce case B because it models what we have seen in Chapter 2 about the behaviour of orienteering data. That is, orienteering data showed a relation between the dispersion of the times and the length of the leg. As the aim is to compare the performance of the asymmetric function we consider data sets with

outliers.

We use the same measures as in Subsection 4.1.1 to compare between the performance of each objective function (asymmetric, least squares, Huber, bisquare) used in the algorithm. Those measures are presented in the following table.

		ASY	LS	HUB	BIS
CASE A	$e_{i,j} \sim N(0,1)$, with outliers				
	a NBIAS	-0.046	0.916	0.630	0.068
	a NMSE	0.002	0.863	0.408	0.005
	b NBIAS	0.001	-0.006	-0.005	-0.001
	b NMSE	0.000	0.001	0.000	0.000
CASE B	$e_{i,j} \sim N(0, b_j^2)$, with outliers				
	a NBIAS	0.016	0.941	0.940	0.025
	a NMSE	0.000	0.917	0.915	0.001
	b NBIAS	-0.001	-0.001	-0.001	0.000
	b NMSE	0.000	0.000	0.000	0.000

ASY: Asymmetric objective function
LS: Least squares objective function
HUB: Huber objective function
BIS: Bisquare objective function

Table 4.4: Results of 1000 simulations to compare objective functions using the algorithm proposed by Maronna and Yohai (2008)

Table 4.4 shows that using an asymmetric function as the loss function in the algorithm proposed by Maronna and Yohai (2008) produces a normalized bias and NMSE of the estimates considerably smaller than those obtained using the Huber or least squares function and a bit smaller than those obtained with the bisquare function. These results show that the use of an asymmetric objective function when the data has skewed outliers seems to produce results as good as the bisquare function and in some cases better.

4.4 Performance of the Improved M&Y algorithm

In Section 4.3 we have presented the results of applying the method proposed by Maronna and Yohai (2008) to produce a lower rank approximation. This section presents the results of applying the Improved M&Y algorithm (our proposed algorithm in Chapter 3). We compared the estimates obtained with the asymmetric objective function against the results obtained using a least squares, a Huber and a bisquare symmetric objective functions.

As we are interested in the effect the errors $e_{i,j}$ and the outliers $u_{i,j}$ have in the estimates, we investigate the following four simulation cases that are an expansion of the cases A and B presented in Section 4.3.

CASE 1: $e_{i,j} \sim N(0,1)$, without outliers.

CASE 2: $e_{i,j} \sim N(0, b_j^2)$, without outliers.

CASE 3: $e_{i,j} \sim N(0,1)$, with outliers.

CASE 4: $e_{i,j} \sim N(0, b_j^2)$, with outliers.

The data used for this analysis were simulated as described in Section 4.1. Similar to Section 4.3 our interest is in the estimates of the vectors, so we follow the methodology in Subsection 4.1.1 to assess the accuracy of the estimates obtained with the Improved M&Y algorithm described in Subsection 3.3.5.

As mentioned in Chapter 2, one objective of implementing this procedure was to use the weights given by the method to identify the outliers in the data. For this reason we would also like to study how accurate it is to use the weights assigned for each objective function as a method to identify outliers.

To do that we define a matrix of optimal weights **W0** based on the outliers added in the simulated data set. This matrix will have only zero and one values. $W0_{i,j} = 1$ if no outlier was added to the element in the i-th row and j-th column, i.e. if $u_{i,j} \times I_{i,j}^{\{0,1\}} = 0$, and $W0_{i,j} = 0$ if an outlier was added to the element in the i-th row and j-th column, i.e. if $u_{i,j} \times I_{i,j}^{\{0,1\}} > 0$. The mean of the differences between the matrix **W0** and the matrix of weights given by the algorithm is a measure of how accurate the weights are detecting true outliers. We calculate the mean and standard deviation of those differences over the 1000 simulations as a measure of the objective function performing the outlier detection.

The results of these measures are used to make a comparison between the performance of each objective function (asymmetric, least squares, Huber, bisquare). The results are presented in Table 4.5.

From the results in Table 4.5 we have that when the data does not have outliers (cases 1 and 2) the least squares, Huber and bisquare objective functions have the lowest normalized bias and mean squared error. But the asymmetric function also produce good results. However when the data has outliers the asymmetric, Huber and bisquare functions perform better than the least squares, and the asymmetric is the one with lowest values of normalized bias and NMSE for the estimated vectors **a** and **b**. As for the identification of the outliers through the weights the asymmetric, Huber and bisquare functions perform very similarly.

We can compare cases A and B from Table 4.4 with cases 3 and 4 from Table 4.5, they have the same simulation parameters and objective functions, the difference

		ASY	LS	HUB	BIS
CASE 1	$e_{i,j} \sim N(0,1)$, without outliers				
	a NBIAS	-0.012	0.000	0.000	0.000
	a NMSE	0.000	0.000	0.000	0.000
	b NBIAS	-0.001	-0.001	-0.001	-0.001
	b NMSE	0.000	0.000	0.000	0.000
CASE 2	$e_{i,j} \sim N(0, b_j^2)$, without outliers				
	a NBIAS	-0.003	0.000	0.000	0.000
	a NMSE	0.000	0.000	0.000	0.000
	b NBIAS	0.000	0.000	0.000	0.000
	b NMSE	0.000	0.000	0.000	0.000
CASE 3	$e_{i,j} \sim N(0,1)$, with outliers				
	a NBIAS	0.016	0.967	0.209	0.018
	a NMSE	0.000	0.965	0.045	0.000
	b NBIAS	0.001	-0.001	0.003	0.000
	b NMSE	0.000	0.000	0.000	0.000
	weight bias	0.001	-0.200	-0.024	0.030
	weight sd	0.004	0.000	0.005	0.003
CASE 4	$e_{i,j} \sim N(0, b_j^2)$, with outliers				
	a NBIAS	-0.002	0.957	0.080	0.008
	a NMSE	0.000	0.948	0.007	0.000
	b NBIAS	0.000	0.000	0.000	0.000
	b NMSE	0.000	0.000	0.000	0.000
	weight bias	0.011	-0.200	0.013	0.035
	weight sd	0.001	0.000	0.009	0.003

ASY: Asymmetric objective function with the Improved M&Y algorithm
LS: Least squares objective function with the Improved M&Y algorithm
HUB: Huber objective function with the Improved M&Y algorithm
BIS: Bisquare objective function with the Improved M&Y algorithm

Table 4.5: Results of 1000 simulations to compare objective functions.

is the algorithm used. From this comparison we see that the Improved M&Y algorithm improves the estimates of the Huber and bisquare function from the estimates obtained with the original algorithm proposed by Maronna and Yohai (2008).

To have a better understanding of the performance of each objective function, we analysed in detail the results of only one simulation for each case. First we present a boxplot of the simulated data. Then we analyse the effectiveness of the objective functions through the residuals and the vector estimates. For the cases with outliers we also study the accuracy of the outlier detection for each function.

4.4.1 Case 1: $e_{i,j} \sim N(0,1)$, without outliers

This first case assumes that the error has a standard normal distribution. As can be seen in Figure 4.7 the dispersion of the observed $t_{i,j}$ is fairly similar across the columns as will be expected for data with identically distributed errors.

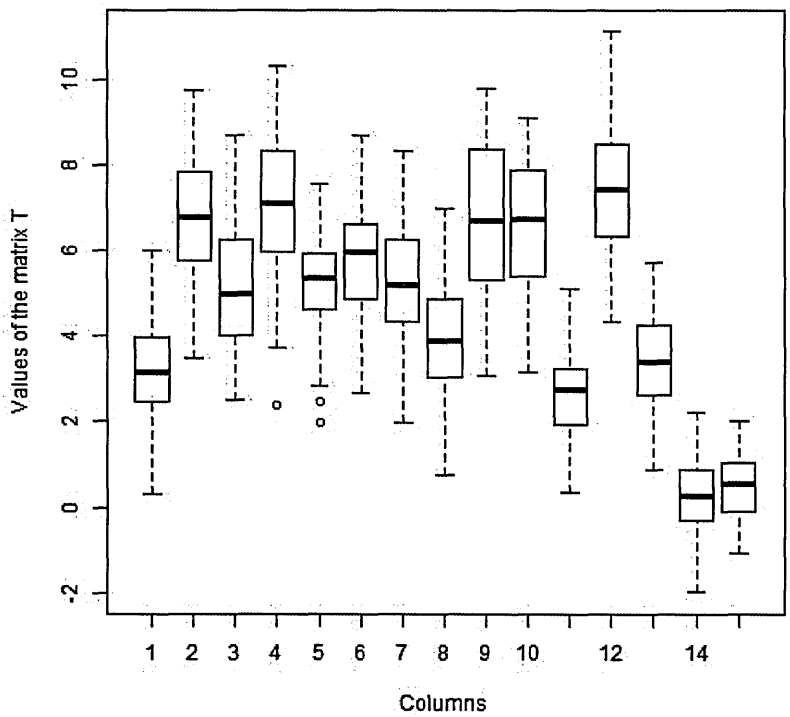
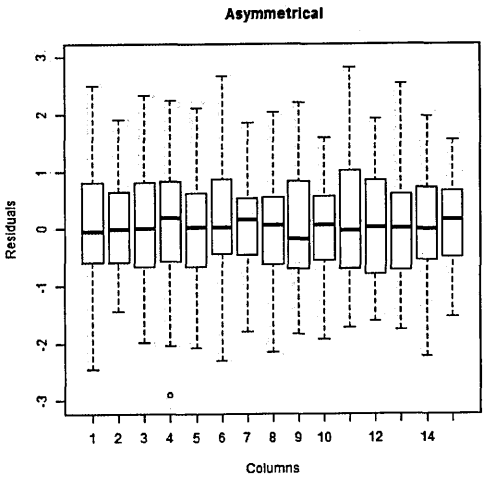


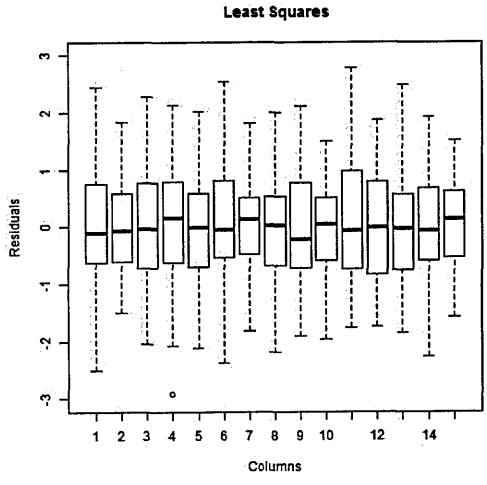
Figure 4.7: Distribution of the simulated data per column.

The residuals $r_{i,j} = t_{i,j} - a_i b_j$ are plotted by column in Figure 4.8 to show how the estimated values fit the real values. For a good fit we will expect the residuals to be zero or near to zero. The plots show the different residuals obtained for each objective function tested. When the algorithm is working well, the residuals should behave similarly to the original errors plotted on Figure 4.8e.

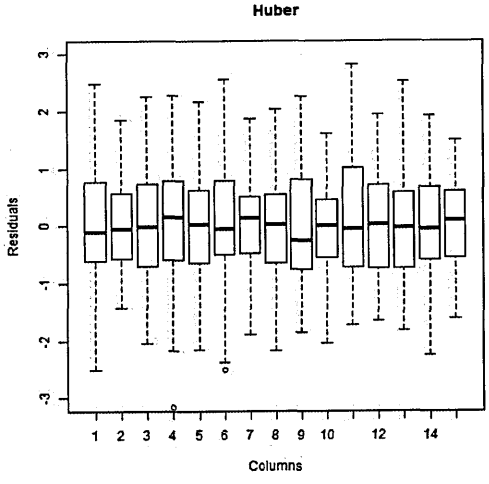
There is not enough evidence to suggest that any of the objective functions perform better than the rest because the plots are similar. All the residuals are significantly close to zero in this case, suggesting a good fit of the estimates.



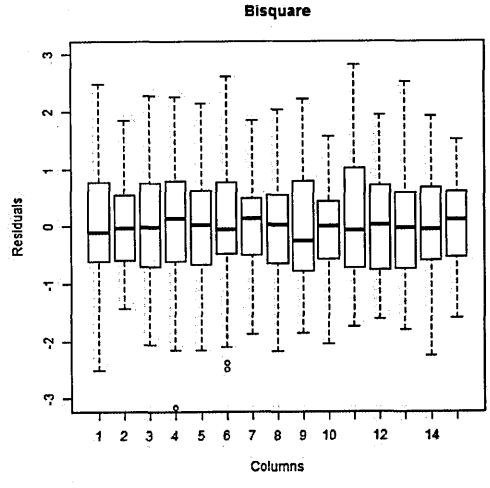
(a)



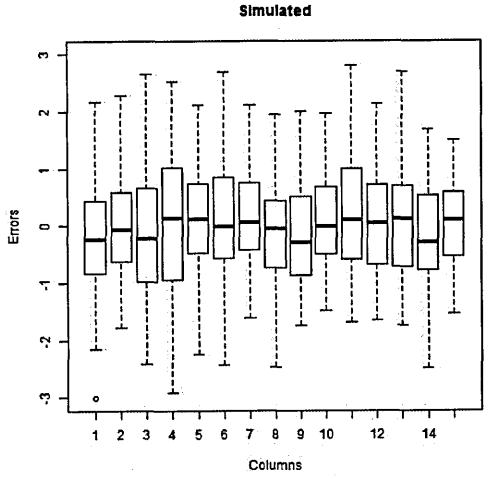
(b)



(c)



(d)



(e)

Figure 4.8: Distribution of the residuals per column for each objective function.

Besides small residuals we also want the estimated vectors to be as near as possible to the original values. The following plots given in Figure 4.9 compare the estimates and the original values for the vector **a**. A similar analysis was done for vector **b**, obtaining similar results.

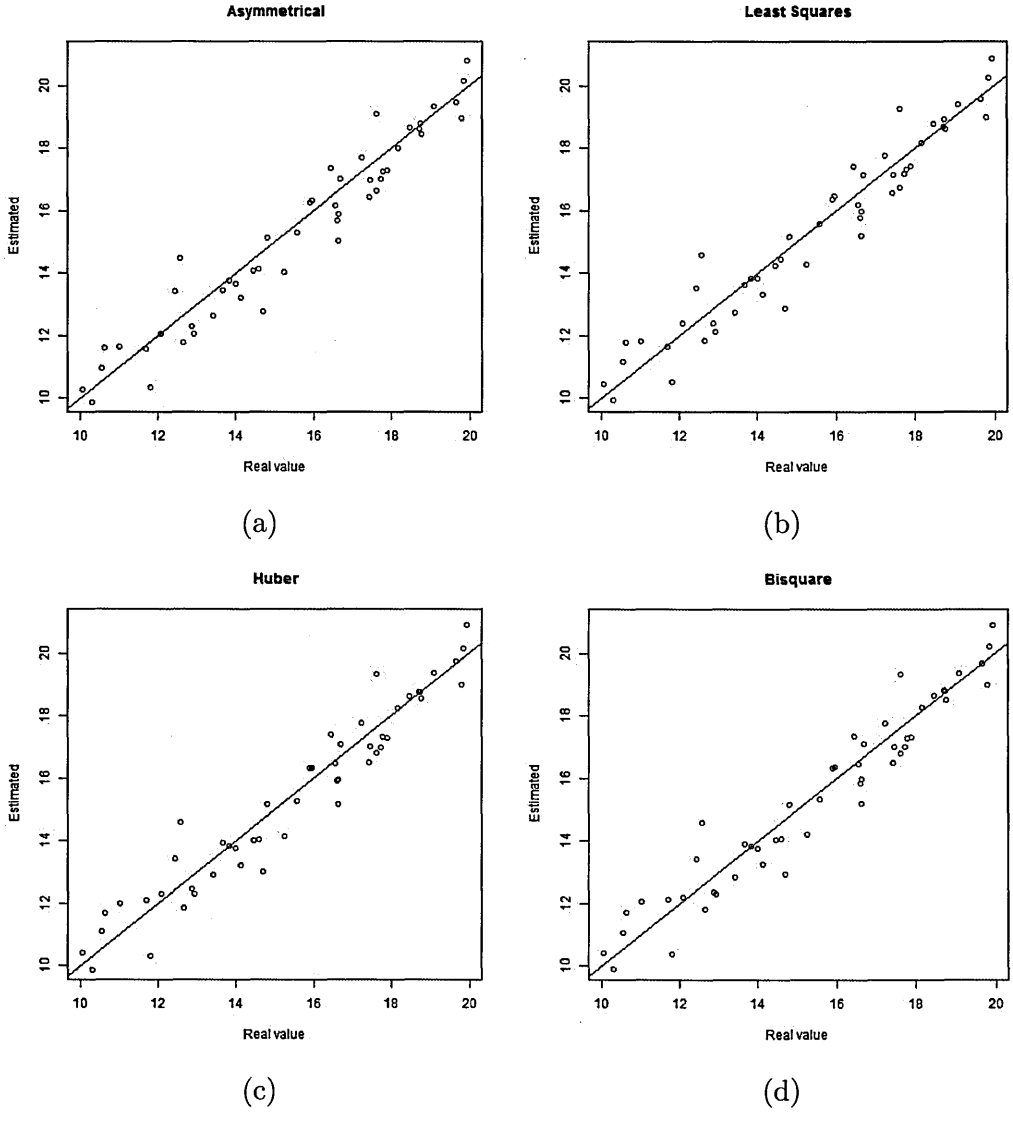


Figure 4.9: Comparison between estimated and original values of the vector **a** for each objective function. The line plotted is $y = x$.

All the functions produced good estimates of the vectors, because all the points are close to the identity function (the diagonal line added in each plot). Considering the results for the residuals and the estimates we conclude that for this case, where the errors have a standard normal distribution, and there are no outliers on the data; all

the objective functions perform similarly producing good estimates of the vectors **a** and **b**.

4.4.2 Case 2: $e_{i,j} \sim N(0, b_j^2)$, without outliers

For this case the assumption is that the dispersion parameter of the error distribution is not constant. Furthermore, because $e_{i,j} \sim N(0, b_j^2)$, the standard deviation of the error distribution will be different for each column and dependent on the vector **b**. Figure 4.10 shows the distribution of the data grouped by column in our simulated data set.

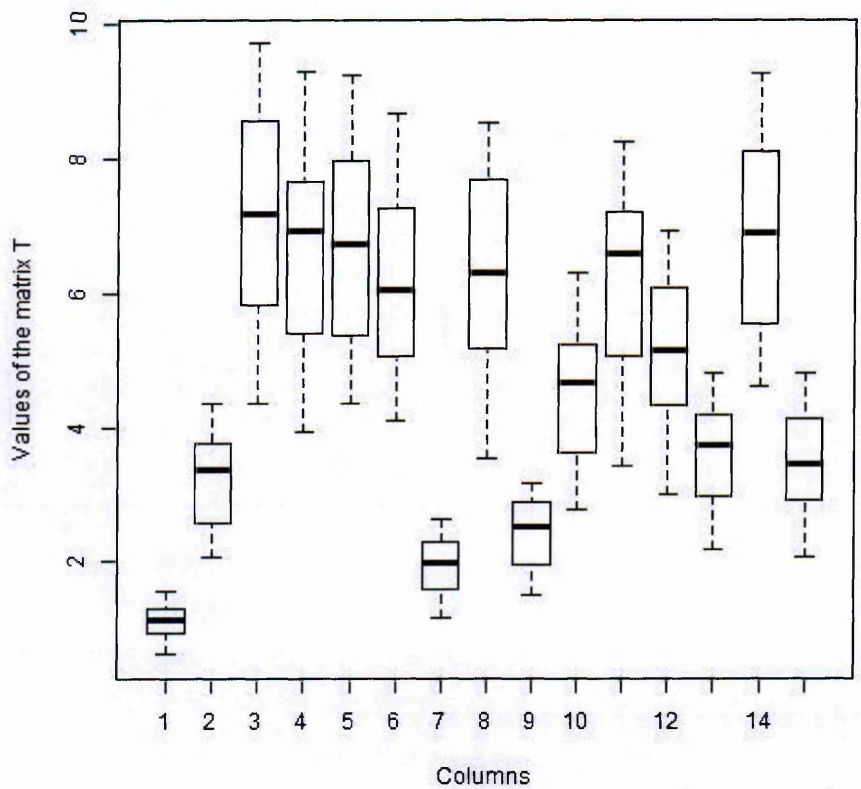


Figure 4.10: Distribution of the simulated data per column.

In this figure we can see the different dispersion values across the legs. For example leg 1 has a very small dispersion, having all its observations near to the median, on

the other hand leg 3 has a larger dispersion having observations that go from over 4 up to around 10 units.

Figure 4.11 shows the plots of the residuals obtained for each objective function. Similarly to case 1, we say that the procedure generates a good fit if the residuals behave similar to the original errors used to generate the data set. Those original errors are plotted in Figure 4.11e.

The plots show that there is not enough evidence to suggest that any of the objective functions performs better than the rest. All the residuals are near zero suggesting a good fit of the estimates.

We analyse the goodness of fit of the method by comparing the estimates and the original values for both vectors. The results are similar for both vectors so we only plot the results for the vector **a**. Figure 4.12 suggests that the vector estimates are close to the real values as the points in the plot lie on or near the diagonal line that represents the identity function.

Considering the results in the residuals and the estimates we conclude that for this case, where the errors have a normal distribution with the dispersion parameter in function of the columns and without outliers on the data, all the objective functions perform similarly producing good estimates of the vectors **a** and **b**.

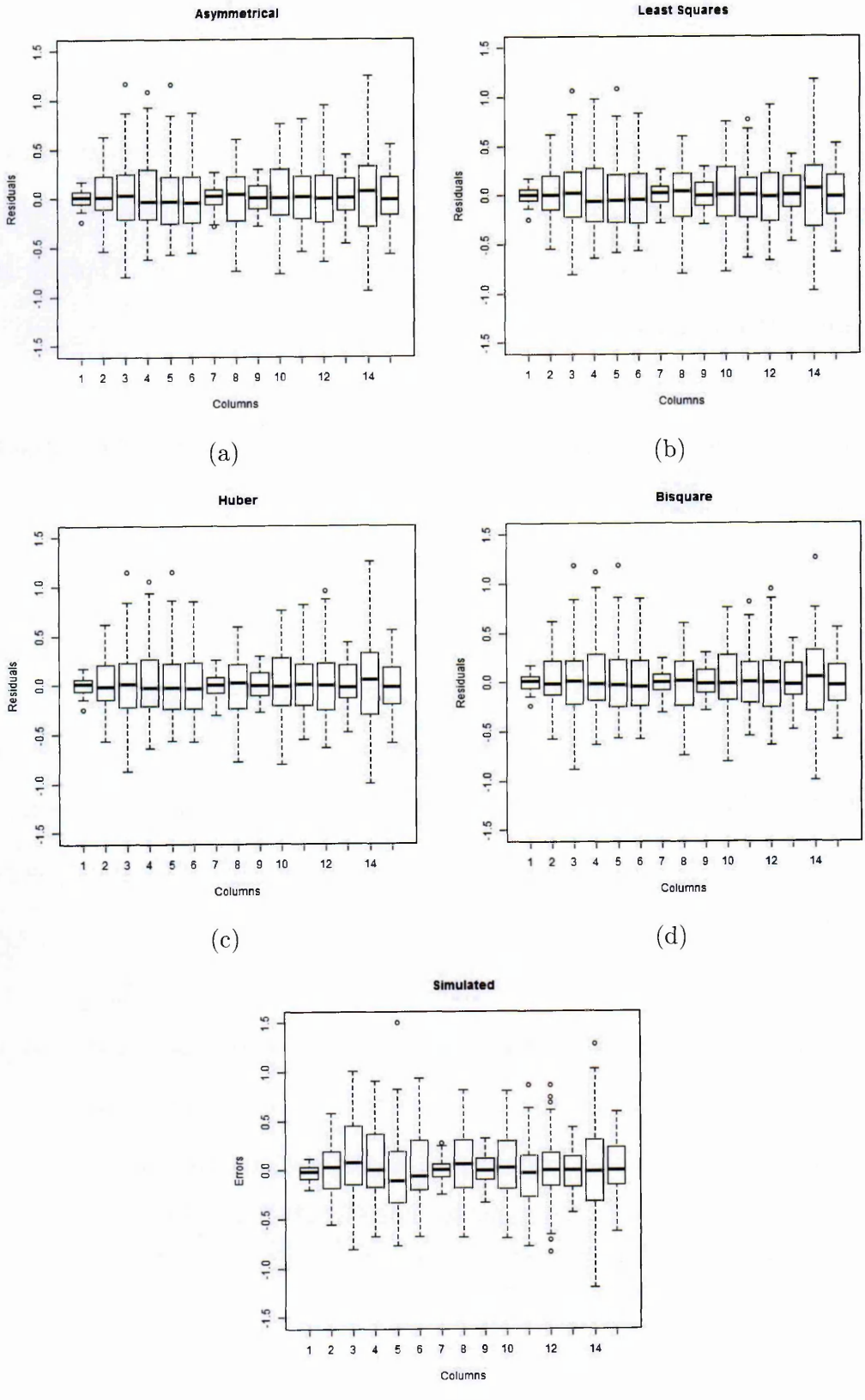


Figure 4.11: Distribution of the residuals per column for each objective function.

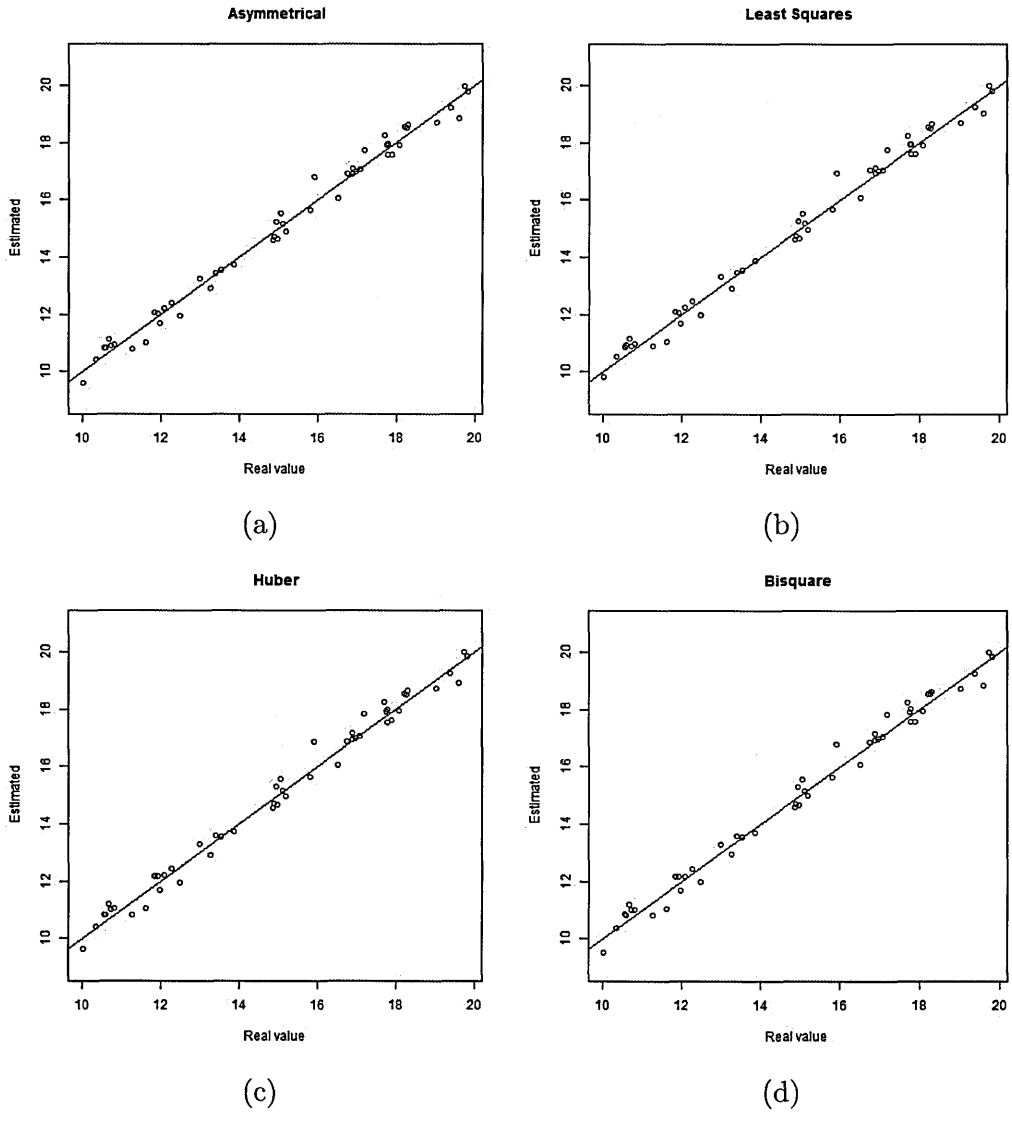


Figure 4.12: Comparison between estimated and original values of the vector **a** for each objective function. The line plotted is $y = x$.

4.4.3 Case 3: $e_{i,j} \sim N(0,1)$, with outliers.

This case uses errors with standard normal distribution as in case 1, but now outliers are added to the data. Figure 4.13 shows the outliers in the data as the points outside the whiskers of the boxplots. The similar dispersion of the observations on each leg caused by the error having a standard normal distribution can also be appreciated, a similar pattern to what was shown in Figure 4.7 .

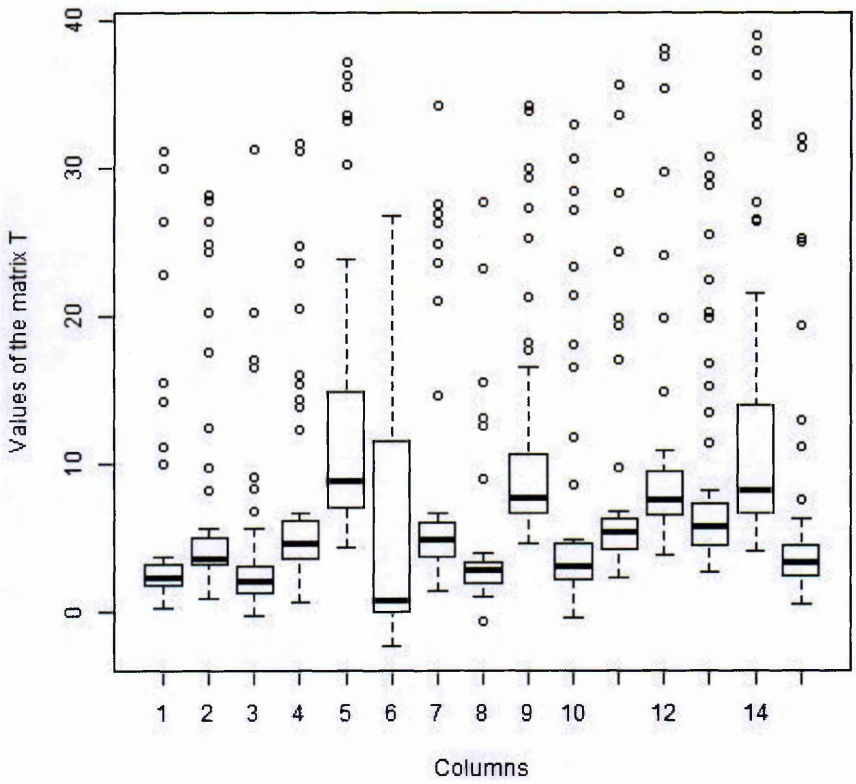


Figure 4.13: Distribution of the simulated data per column, including the outliers.

It is important to mention that when outliers are added it becomes difficult to differentiate between the error dispersion and an outlier, specially when using the boxplot as identification method. Going back to the simulated data is possible to verify that columns 5 and 6 have 14 outliers each and column 14 has 13 outliers. On the other hand column 8 only has 6 outliers. Figure 4.13 shows that those columns with the largest number of outliers are the ones with the largest dispersions of their observations. This means that for those columns the boxplot method is not detecting all the outliers.

Figure 4.14 presents the plots of the residuals obtained for each objective function tested. When the method approximates the real data it is expected that the residuals behave similarly to the original errors and outliers plotted on the last graph (e).

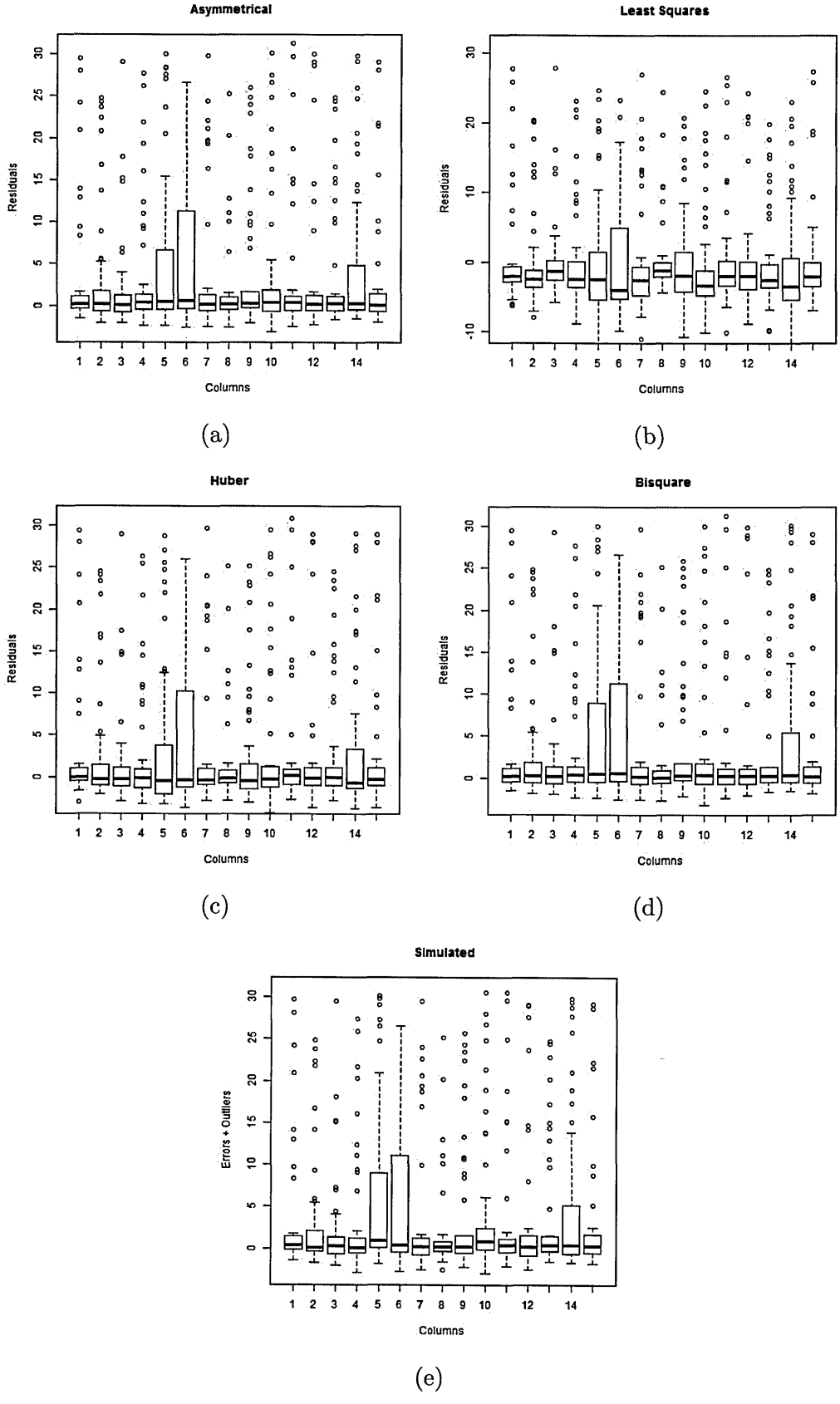


Figure 4.14: Distribution of the residuals per column for each objective function.

We can see that the residuals for the least squares objective function are pulled to the negative values. This effect in the residuals is caused by the fact that the least squares function is not robust to the presence of outliers. For the other functions the plots of the residuals do not seem to differ much, and they look similar to the boxplot of the errors plus outliers used on the original data set in Figure 4.14e. So for this case we have that the residuals suggest the selection of an objective function such as asymmetric, Huber or bisquare produces better estimates. However these plots do not clearly identify which of these functions is the best.

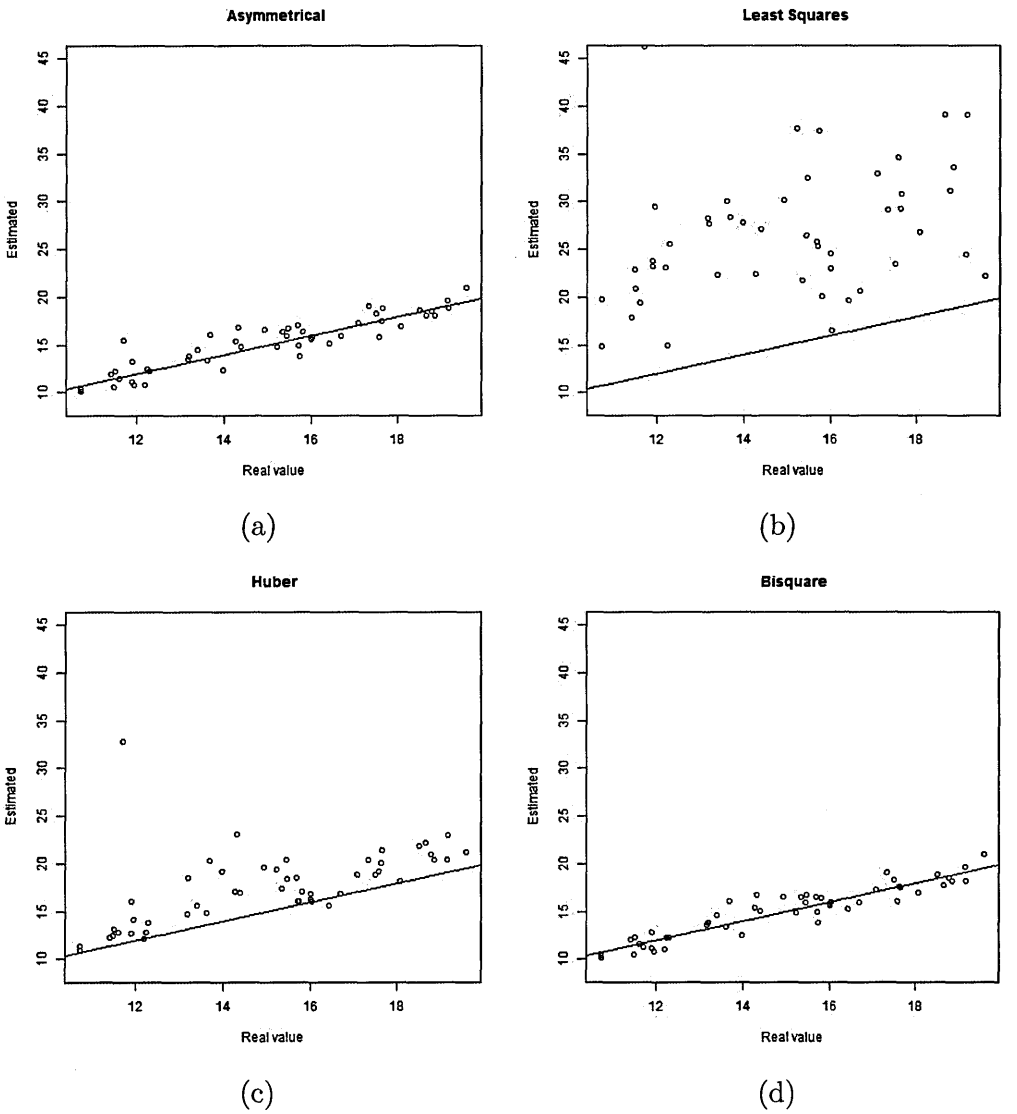


Figure 4.15: Comparison between estimated and original values of the vector **a** for each objective function. The line plotted is $y = x$.

The goodness of fit of the method for a single data set is analysed by comparing the estimated and the original values for both vectors. Figure 4.15 presents the results for the vector **a**. The results for the other vector are similar.

Figure 4.15 clearly shows that for the least squares and Huber functions the points in the plots are above the identity line. This suggests that the vectors are being over estimated, this is a consequence of the presence of positive outliers. However the asymmetric and bisquare functions produce estimates close to the real values.

We want our method to be very efficient also in the identification of outliers. So we investigated how these functions coped with the identification of outliers. We plotted in Figure 4.16 the weights against the outliers simulated and added to the data set. The vertical line in the plots represents the cut point for outliers according to the parameters used in the simulation, so all the points on the right side of that line were outliers on the original data.

The least squares function by definition assigns weights equal to one to all the observations, so it is impossible to distinguish the outliers through the weights. The Huber function seems to miss a few outliers which the bisquare and asymmetric functions do not. The plots show that there is a trade off between the value of the weights and the definition of outlier using this values. For the asymmetric function this point is near 0.8. So if the weight is lower than that almost certainly it is an outlier, but for the bisquare this point is around 0.6. These cut points suggest that the asymmetric function is more effective at ignoring the outliers on the estimates calculations, which might explain the effect on the normalized bias and NMSE values presented in Table 4.5.

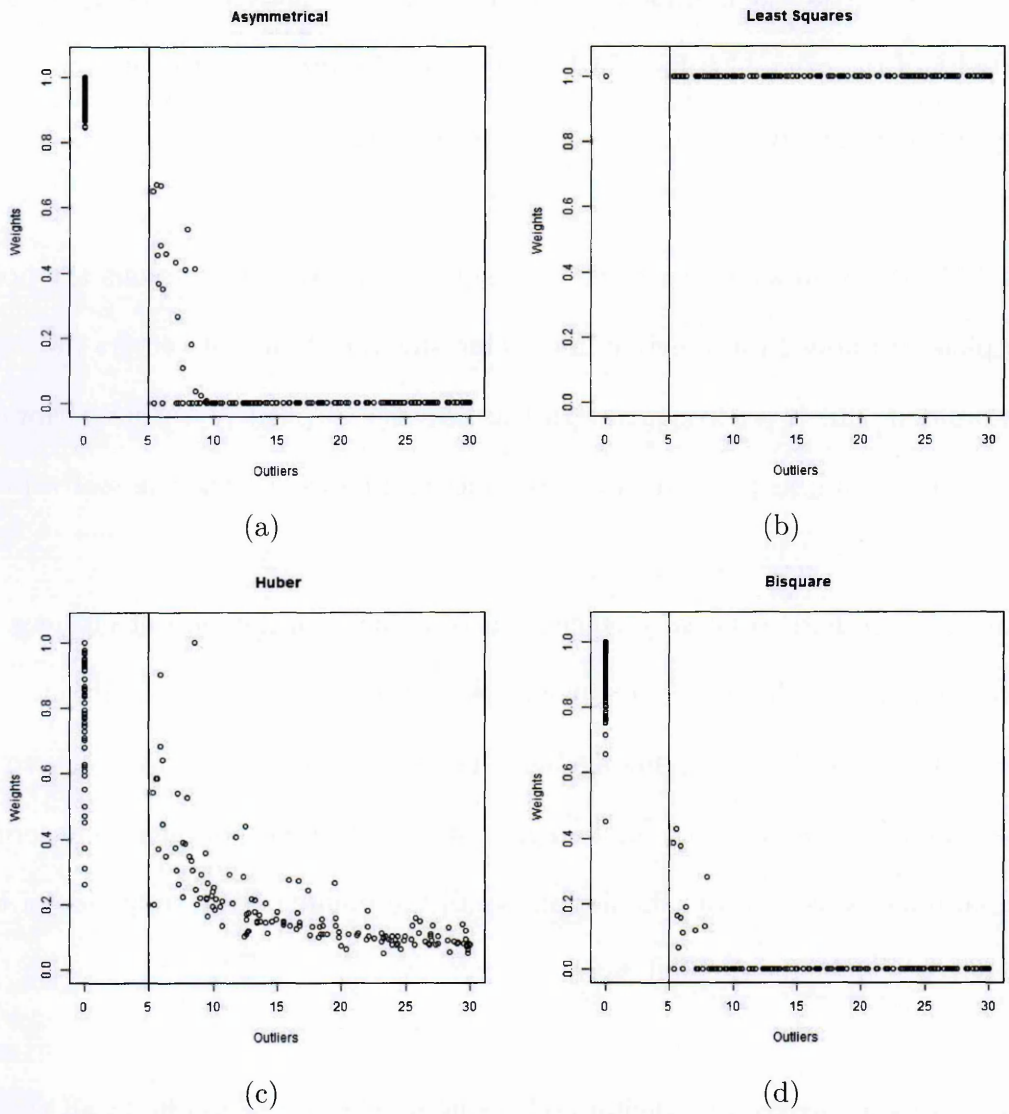


Figure 4.16: Comparison between outliers and the weights for each objective function.

4.4.4 Case 4: $e_{i,j} \sim N(0, b_j^2)$, with outliers

This case is the closest to what we think is the behaviour of orienteering data. The data are simulated with errors that have the standard deviation of the normal distribution dependent on the vector **b** similar to case 2, and outliers are also added.

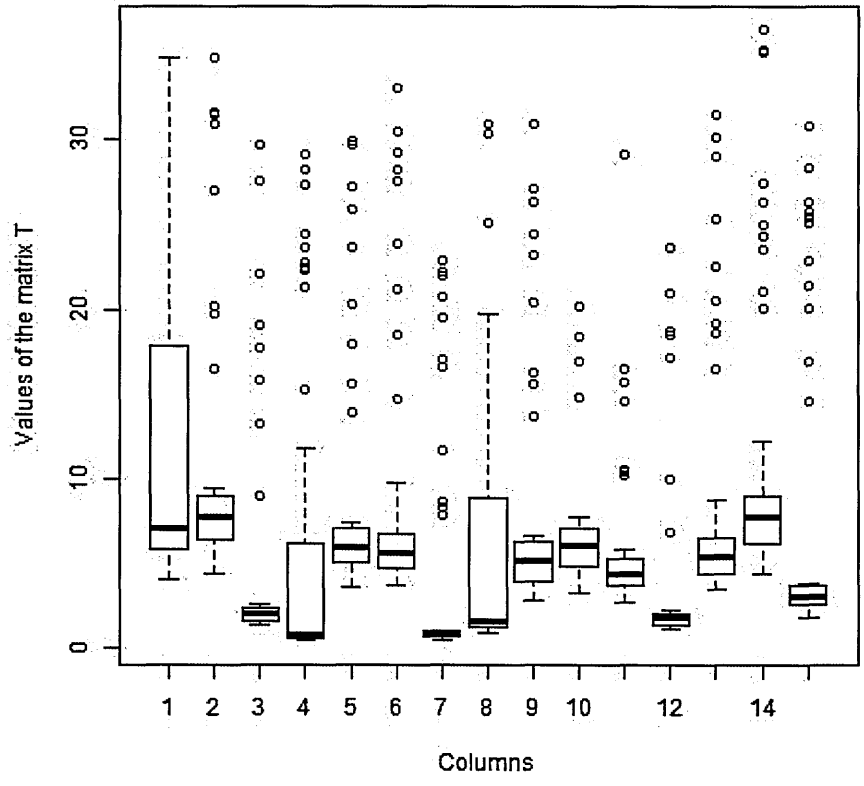
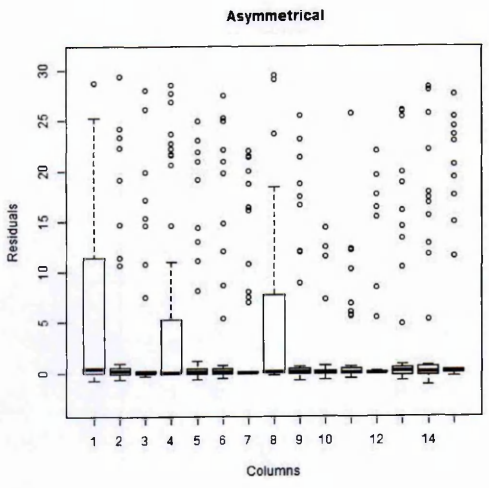
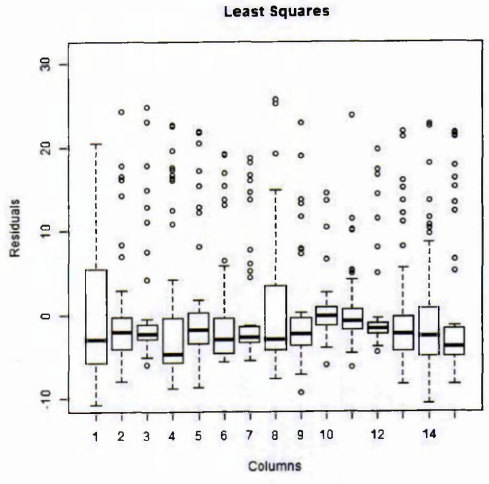


Figure 4.17: Distribution of the simulated data per column, including the outliers.

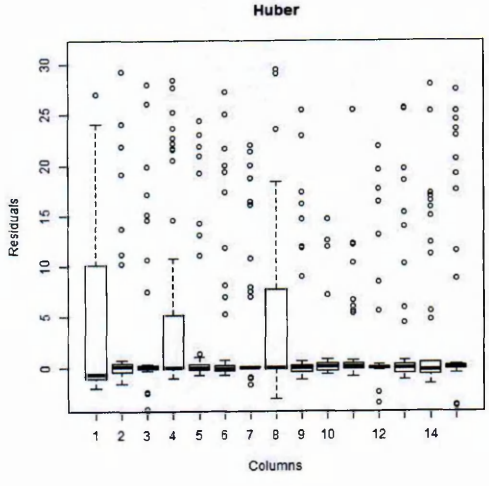
Figure 4.17 shows that in this particular data set the columns with more dispersion are 1, 4 and 8, and this will also appear in the residuals plots in Figure 4.18. Going back to the simulated data it turns out that column 1 has 18 outliers and columns 4 and 8 have 13 outliers each. On the other hand column 10 has only 4 outliers and the columns 3, 7 and 12 have smaller dispersion. The data indicates that columns 3, 7 and 12 have 8, 11 and 7 outliers respectively. Figure 4.17 shows that those columns with the largest number of outliers tend to be the ones which boxplots have the largest boxes. This suggests that the differentiation between the error dispersion and an outlier might be more difficult.



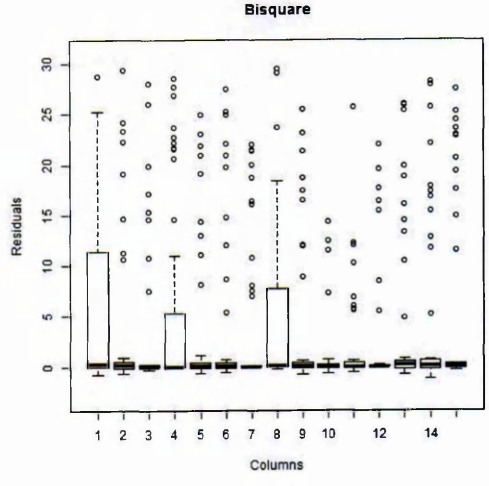
(a)



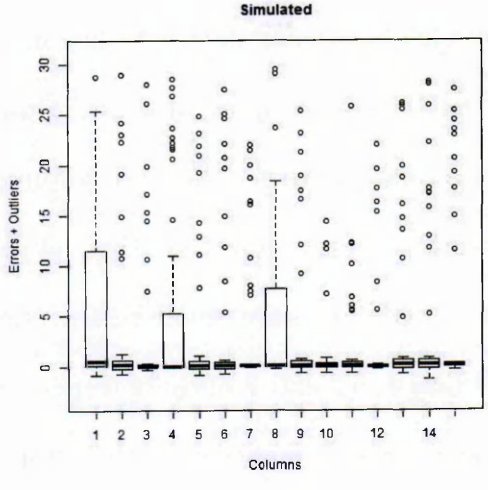
(b)



(c)



(d)



(e)

Figure 4.18: Distribution of the residuals per column for each objective function.

Figure 4.18 shows the plots of the residuals obtained for each objective function. For a reliable fit, it is expected that the residuals behave similarly to the original errors and outliers generated for the simulation, which are plotted on the last graph (e).

In Figure 4.18 we can see that because the least squares function is not robust to the presence of outliers, the residuals for this function are pulled to the negative values and do not have median zero. For the asymmetric, Huber and bisquare functions the plots of the residuals do not differ from each other, and they look similar to the boxplot of the errors plus outliers used on the original data set Figure 4.18e. This suggest that the asymmetric, Huber and bisquare functions perform better than the least squares in this case.

We analyse the goodness of fit of the method by comparing the estimated and the original values for both vectors. The results are similar for both vectors so we only plot the results for the vector **a**.

Figure 4.19 shows that the least squares function is over estimating the vectors. This can be seen in the plots as the points are above the identity line. The Huber function also shows some over estimation, in particular for the three observations furthest away from the identity line. The over estimation of the Huber function is not as pronounced as for the least squares. The plots for the asymmetric and bisquare functions show the points closer to the diagonal line, suggesting that their estimates are nearer to the real values.

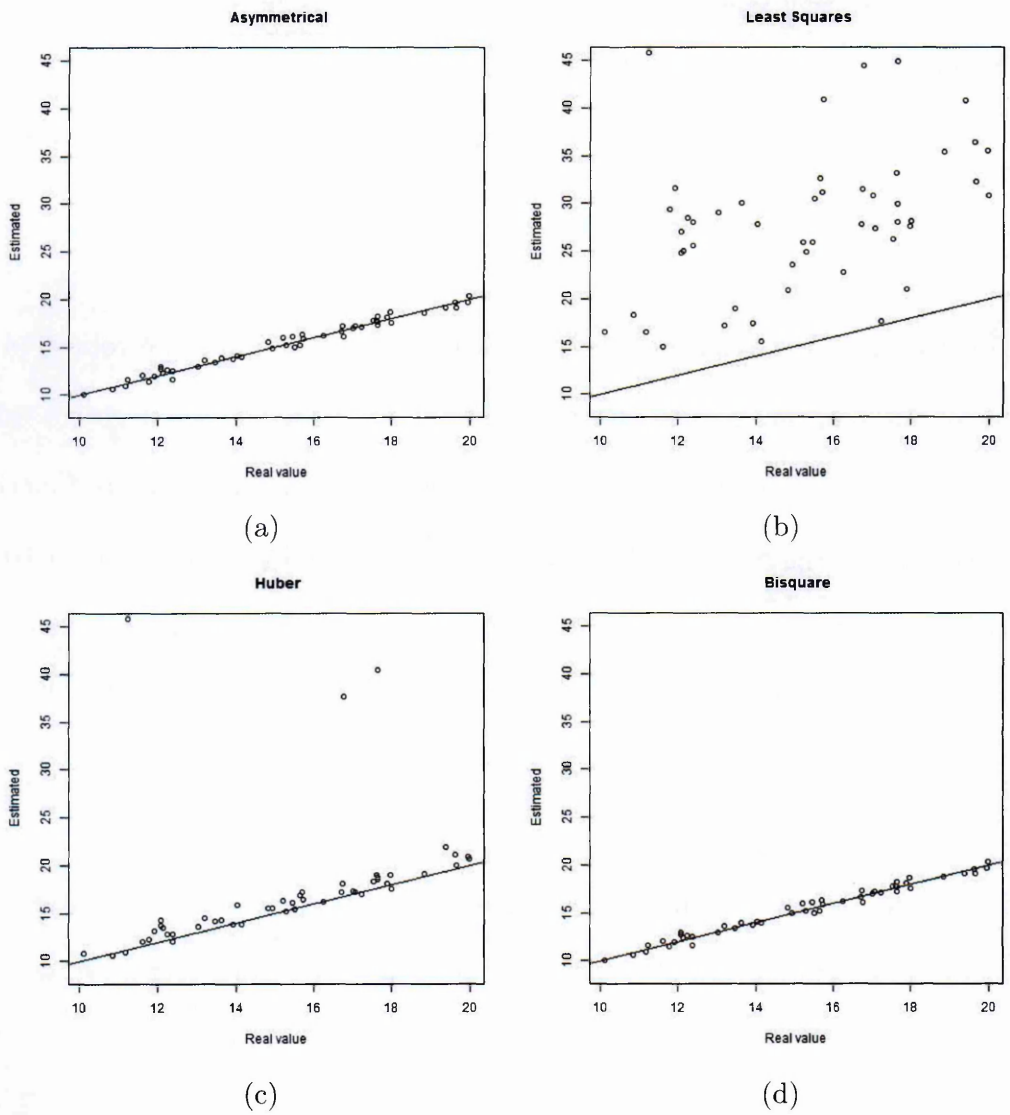


Figure 4.19: Comparison between estimated and original values of the vector **a** for each objective function. The line plotted is $y = x$.

Again we want to investigate how these functions cope with the identification of outliers. In Figure 4.20 we plot the weights against the outliers simulated and added to the data set. The vertical line in the plots represents the cut point for outliers according to the parameters used in the simulation, so all the points on the right side of that line were simulated outliers.

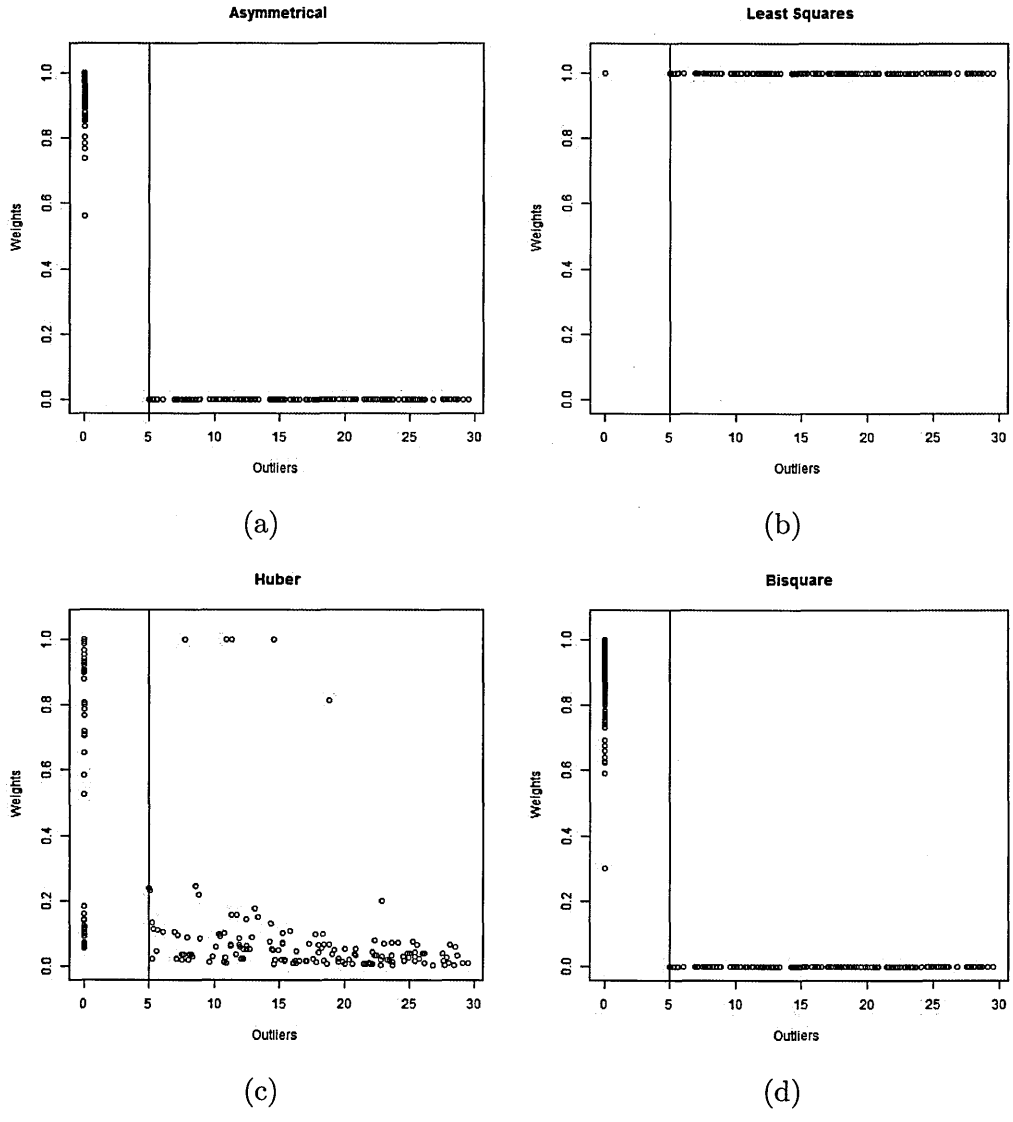


Figure 4.20: Comparison between outliers and the weights for each objective function.

As mentioned before it is impossible to distinguish the outliers through the weights for the case of the least squares objective function. Similar to case 3, the bisquare and asymmetric functions show a clearer identification of outliers in comparison with the Huber function. The plots show that there is a trade off between the value of the weights and the definition of outlier using this values. For the asymmetric function this point is near 0.6 so if the weight is lower than that almost certainly it is an outlier, and for the bisquare this point is around 0.3. These cut points suggest that the asymmetric function is more effective at ignoring the outliers in the

estimates calculations, which might explain the effect on the normalized bias and NMSE values presented in Table 4.5.

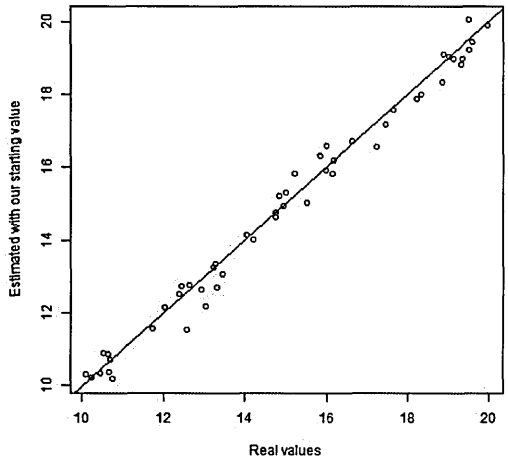
Based on one data set from this simulation study is possible to conclude that for the case of having a data set with normally distributed errors and asymmetric outliers the asymmetric and bisquare functions produce estimates closer to the true values of the vectors **a** and **b**. Also the weights assigned by both functions seem to be accurately detecting the outliers in the data set. And an examination of 1000 simulated data sets, results of which were presented in Table 4.5 show that the asymmetric function has smaller normalized bias and NMSE for the estimates in comparison with the bisquare objective function. This suggests that for type of data sets with asymmetric outliers simulated in this study, the use of an asymmetric objective function produce better estimates than a least squares or Huber function and similar or slightly better estimates than a bisquare objective function.

4.5 Advantages of the Improved M&Y algorithm

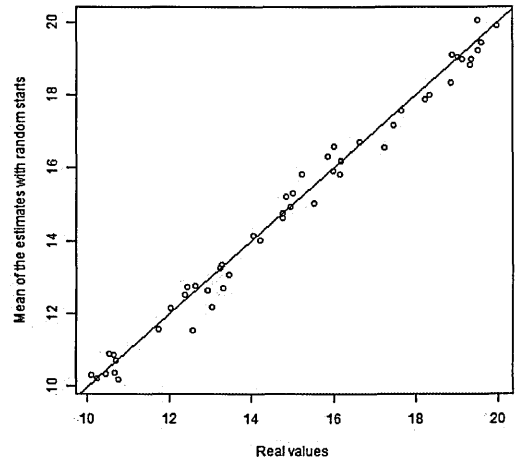
We are interested in analysing in more detail the performance of the Improved M&Y algorithm, but only focusing on data sets that follow the pattern described in case 4. As we have shown in Tables 4.4 and 4.5 the estimates obtained with our proposed algorithm have smaller normalized bias and NMSE than the estimates obtained using the method proposed by Maronna and Yohai (2008). This suggests that the algorithm we propose produces better estimates of the vectors **a** and **b** when the data has asymmetric outliers. We have mentioned that our algorithm is based on modifications to the method proposed by Maronna and Yohai (2008), and we have also said that these modifications provide additional properties for the initial values, the scale parameter and the weights. We study those improvements in this section.

4.5.1 Our method does not need specific initial values

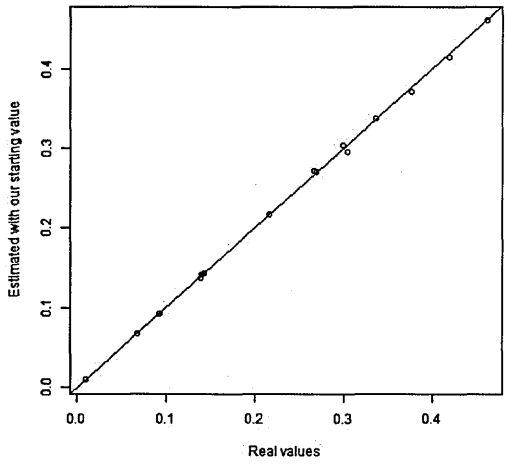
The method proposed by Maronna and Yohai (2008) has a complex procedure to select the initial values due to the sensitivity of their algorithm to the starting point. To study the need for specific initial values to start the the Improved M&Y algorithm we compared the results obtained by initializing the algorithm with our proposed initial values and the results obtained if the algorithm has a random start. Figure 4.21 shows the plots of estimates for both vectors **a** and **b** against their original values used to make the simulation.



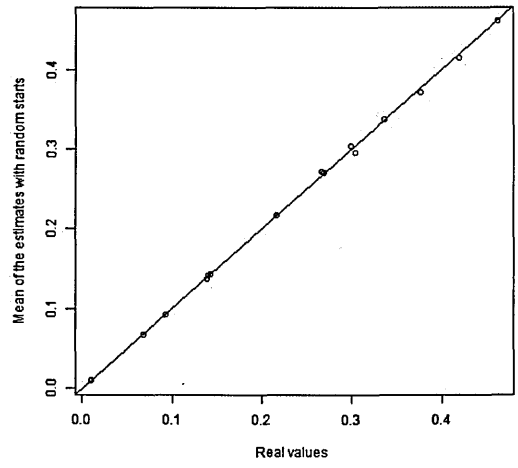
(a) Our initial value for the vector **a**



(b) Random initial value for the vector **a**



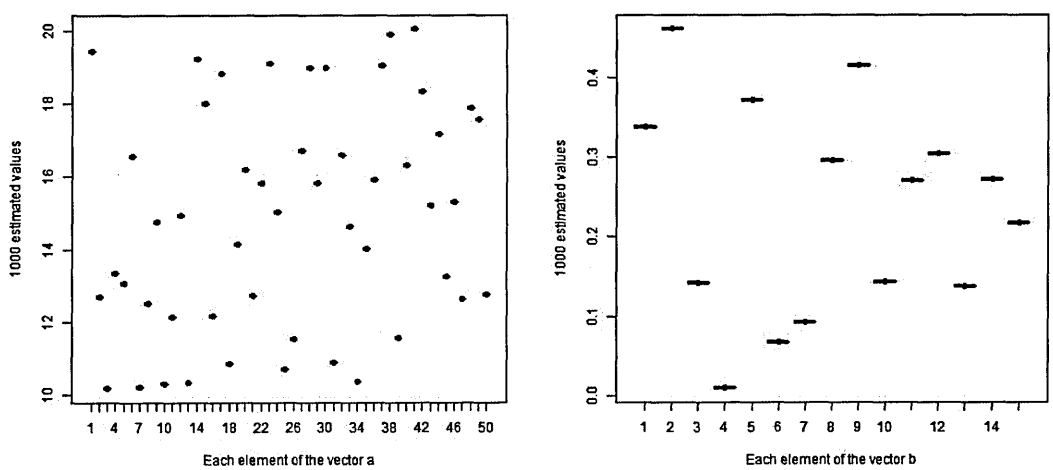
(c) Our initial value for the vector **b**



(d) Random initial value for the vector **b**

Figure 4.21: Estimated values versus the real values for both vectors. The line plotted is $y = x$.

When the initial values proposed by the Improved M&Y algorithm are used, the values of the estimates are very close to the originals. This can be seen in Figures 4.21a and 4.21c, because the points lie near the diagonal line that represents the identity function. This result was what we expected as we defined the starting point to produce good estimates. However this is only one estimate and we need to analyse the estimates of different starting values. We selected randomly 1000 different starts, for each start an estimate of the vectors **a** and **b** was obtained. In Figures 4.21b and 4.21d we plotted the mean of the estimates of the vectors for those thousand starts against the true values. The figures show that on average the estimates are also very close to the original values.



(a) Distribution of the 1000 estimates of the vector **a**.

(b) Distribution of the 1000 estimates of the vector **b**.

Figure 4.22: Distribution of the 1000 estimates for both vectors.

Figures 4.22a and 4.22b present the boxplots of the 1000 estimates of each value of the vectors **a** and **b** respectively. These plots show that the dispersion between these one thousand estimates is near zero, this suggest that the estimates are the same for all the randomly selected initial values. This is a very important characteristic of the Improved M&Y algorithm because it appears we can start the iterations at any point knowing that the estimates obtained are going to be always the same. This

leads us to conclude that the Improved M&Y algorithm produces estimates for **a** and **b** that appear not to be reliant on choosing a good starting point.

4.5.2 The Improved M&Y algorithm under random starts

Continuing the analysis started in Subsection 4.5.1 we analyse what happens to the loss function, the scale parameter and the weights when the initial values are selected randomly. We use the same data set and random starts simulated for the previous subsection. The objective of the algorithm is to minimise a loss function, so it is of interest to know if the minimal value reached by the algorithm differs when different starting points are used.

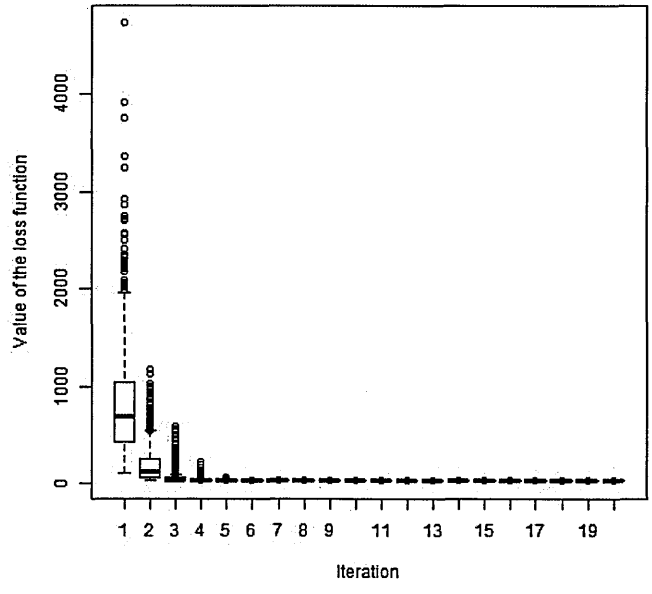


Figure 4.23: Values of the loss function through the iterations.

Figure 4.23 shows that for different starting points, the value of the loss function tends to decrease as the number of iterations increases. This simulation study shows that the final value for the loss function that this algorithm found was the same for all the thousand random starts and it was reached around the 6th iteration. To be certain that the loss function has reached that final value, and because there is little

difference in the processing time between doing six or twenty iterations (each iteration takes on average 0.11 seconds), the algorithm is set to iterate for 20 iterations similarly to one of the stopping rules of the algorithm proposed by Maronna and Yohai (2008).

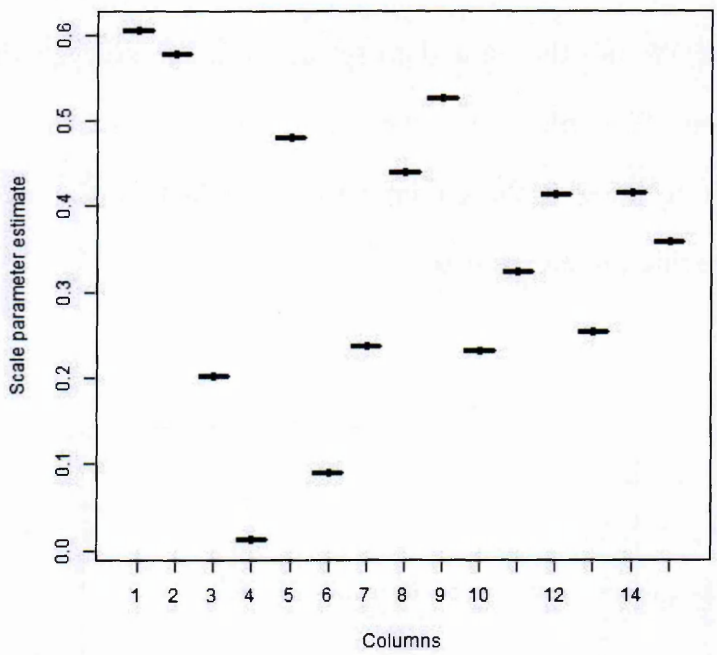


Figure 4.24: Distribution $\hat{\sigma_j}$ after 20 iterations for 1000 different random starts.

In Chapter 3 it was mentioned that our proposed algorithm uses Sn to estimate the scale parameter $\sigma_j (1 \leq j \leq m)$, and this scale estimate is updated in every iteration. Because of this dependence of the scale estimate on the estimates of the vectors **a** and **b**, it is of interest to study the behaviour of this scale estimate when the starting points are selected randomly. Figure 4.24 shows the boxplots of the scale parameters per column. We stated that the scale parameter was column dependent, so what we expect to see in the plot is different scale estimates for each column obtained in each of the 1000 random starts.

Figure 4.24 shows that the boxplots do not indicate any dispersion. This means that the scale estimates by column after 20 iterations are the same for all the simulated random starts. This is a good result because it shows that the parameters in the loss function are reaching a consistent point at the end of the iterations.

In a similar way to the scale parameter, the weights could change from one run to the other. This could happen because the weights work in combination with the estimates of the vectors, so a different combinations of the weights and vectors could give equal estimates of the matrix **T**. The next figure presents the weights corresponding to the first row of the matrix **T** across the 1000 starting points.

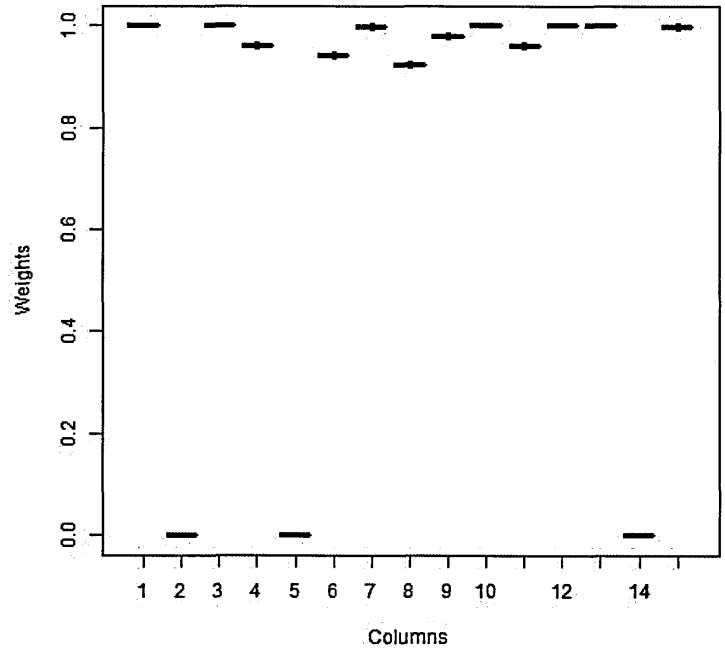


Figure 4.25: Distribution of the weights for the first row after 20 iterations for 1000 different random starts.

Figure 4.25 shows that the weights assigned for the 15 columns of the first row are always the same. Similar plots were obtained for the other 49 rows of the simulation study. This means that the weights given to each element of the matrix by the Improved M&Y algorithm are also independent of the starting point.

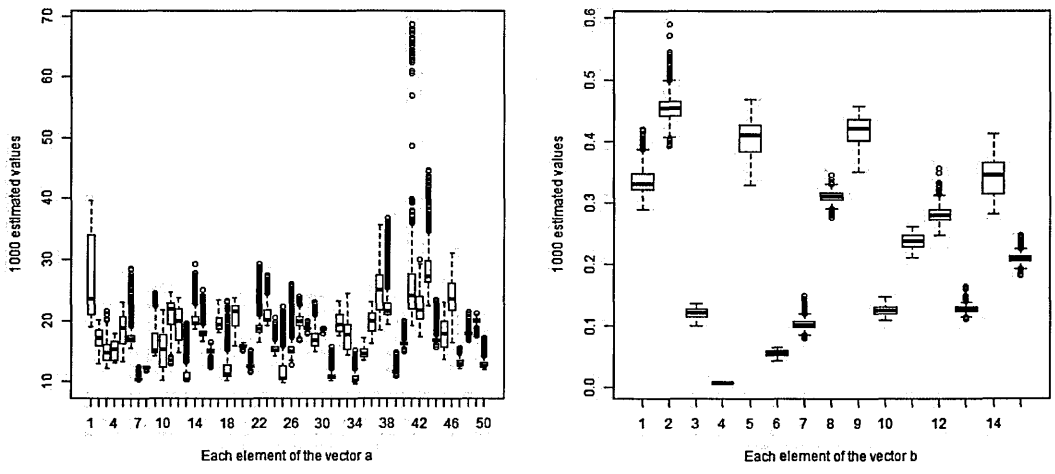
This analysis seems to suggest that for any starting point our proposed algorithm will produce the same results. This means that the performance of the Improved M&Y algorithm does not seem to depend on the initial values.

4.5.3 Maronna and Yohai algorithm under random starts

We have seen how our proposed algorithm performs with random initial values. Now we analyse the effect of random starts on the algorithm proposed by Maronna and Yohai (2008). The data set used is the same as those for the previous subsections, and the 1000 random starts are the same as the ones used to test the Improved M&Y algorithm.

The algorithm proposed by Maronna and Yohai (2008) uses the bisquare objective function, uses the M-scale estimate of the scale parameter, given by equation (3.27) in Subsection 3.3.2 and selects one of the rows as its initial value. In Section 4.3 it was mentioned that this algorithm produces good estimates of the vectors **a** and **b**. However the following results will show that if the initial values are changed the estimates obtained change as well.

Figure 4.26 presents the boxplots of the 1000 estimates for each vector **a** and **b**. The dispersion observed in the plots in Figures 4.26a and 4.26b suggests the vector estimates obtained through the algorithm differ between random starts. In contrast to the Improved M&Y algorithm, the estimates produced by the algorithm proposed by Maronna and Yohai (2008) are dependent on the initial values.



(a) Distribution of the 1000 estimates of the vector **a**.

(b) Distribution of the 1000 estimates of the vector **b**.

Figure 4.26: Distribution of the estimates for both vectors based on the same data set with 1000 different random starts.

As mentioned in Chapter 3 this algorithm estimates the scale parameter using the initial vectors and then is fixed through the iterative process. For this reason if random starts are used then the scale parameter estimate will be different in each case. Figure 4.27 presents the boxplot of the scale parameter estimate over the 1000 simulated random starts, and is possible to observe that the estimates are different for each starting point as was expected to happen. This results seem to show that the sensitivity of the estimates for the vectors **a** and **b** to the starting point might be a result of the scale parameter not being updated.

In the previous subsection we analysed how the weights are assigned for the different starts with the Improved M&Y algorithm. We will now study the weights under random starts using the algorithm proposed by Maronna and Yohai (2008). Given that results for the estimates of the vectors **a** and **b** and $\hat{\sigma}_j$, we expect the weights will be also dependent on the starting point.

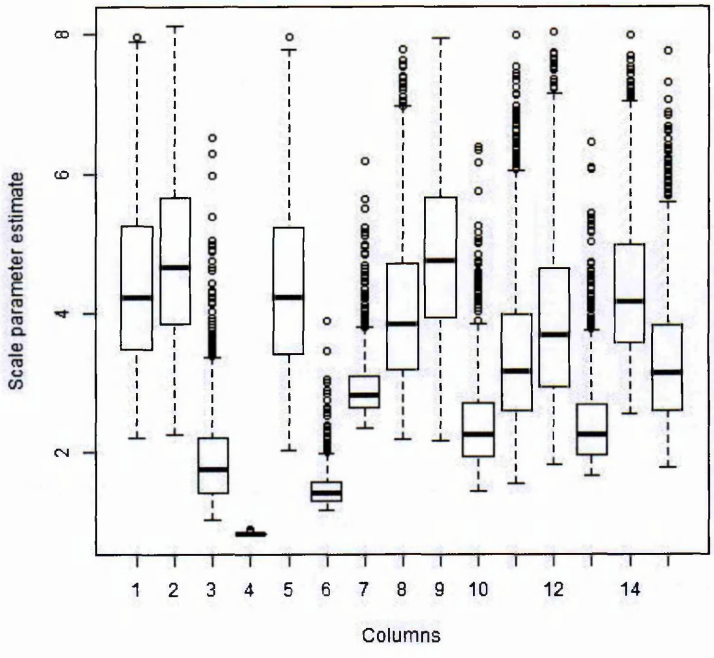


Figure 4.27: Distribution $\hat{\sigma_j}$ after applying the algorithm for 1000 different random starts.

Figure 4.28 presents the weights assigned to the 15 columns of the first row for the 1000 different starting points. Similar plots were obtained for the other 49 rows of the simulation study. The dispersion observed in all the box plots suggests the weights assigned by the algorithm for a particular observation could be different depending on the starting point of the iterative process. For example column 14 on this first row of has weights that go from zero to over 0.9. This suggest that, like the estimates of the vectors, the weights given to each element of the matrix by the algorithm proposed by Maronna and Yohai (2008) are also dependent on the starting point.

In conclusion the algorithm proposed by Maronna and Yohai (2008) does not perform properly if the initial values are chosen randomly and this agrees with the observation made by Maronna and Yohai (2008) relating to the importance of hav-

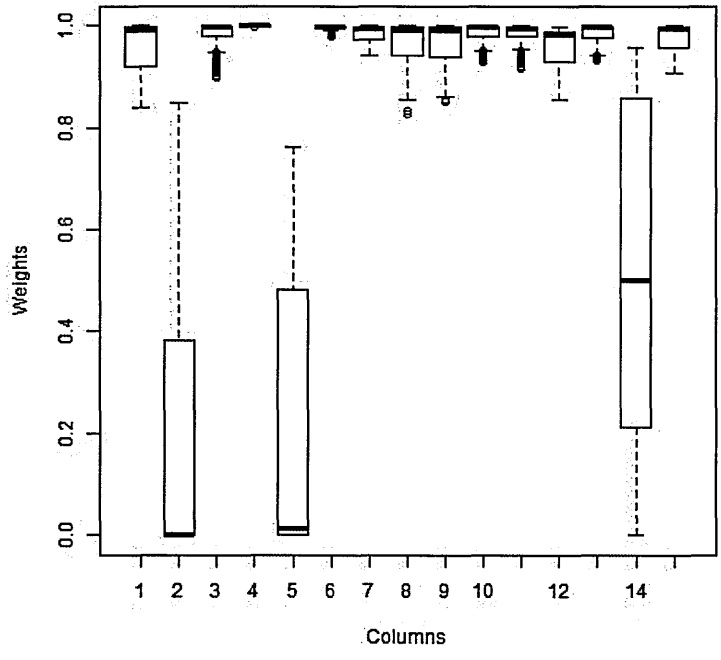


Figure 4.28: Distribution of the weights for the first row after applying the algorithm for 1000 different random starts.

ing a robust and appropriate initial values estimate in order for the algorithm to perform correctly.

4.5.4 Loss function decreasing at each iteration

In Chapter 3 we mentioned that the proof of convergence of the loss function to a minimum depends on non-updating the scale parameter at each iteration. We also stated that when the scale parameter is updated at each step (as we proposed in our modified algorithm), the point reached by the loss function might not be the minimum, but it corresponds to estimates as good as the ones obtained with the algorithm proposed by Maronna and Yohai (2008).

Table 4.6 presents a comparison between the Improved M&Y algorithm (our proposal) and the algorithm proposed by Maronna and Yohai (2008) using the bisquare and asymmetric objective functions. We also tested a semi-improved algorithm,

which uses Sn as the scale parameter but is not updated at each iteration as in the Improved M&Y algorithm. Because the scale parameter is not updated the selection of initial values becomes important in the semi-Improved M&Y algorithm. We use the initial values described in Subsection 3.3.1.

The following table presents the normalized bias (NBIAS) and normalized mean squared error (NMSE) of the vectors **a** and **b** estimates. For this analysis we simulated 1000 data sets following the distributions described in Section 4.1 and case 4 in Section 4.4.

Function	BIS	ASY	ASY	ASY
Algorithm	M&Y	M&Y	semi-Improved M&Y	Improved M&Y
a NBIAS	0.025	0.012	-0.001	-0.002
a NMSE	0.001	0.000	0.000	0.000
b NBIAS	-0.016	-0.053	-0.005	-0.003
b NMSE	0.004	0.044	0.000	0.000
weight bias	-0.004	0.005	0.008	0.011
weight sd	0.002	0.018	0.002	0.001

Table 4.6: Results of 1000 simulations to compare different algorithms

Table 4.6 shows that an improvement in the estimates occur when the asymmetric function is introduced into the algorithm proposed by Maronna and Yohai (2008). However a larger improvement in the normalized bias and mean square error is observed when the improvements to the algorithm proposed by Maronna and Yohai (2008) are implemented. The results show a small difference between the Improved M&Y algorithm and the semi-Improved M&Y, which means that in terms of the normalized bias and mean square error there is little to gain by updating Sn at every iteration. This suggests that the semi-Improved M&Y algorithm gives the option of an algorithm which loss function decreases at each iteration, however, we will inherit the need for appropriate initial values for good estimates and weights.

Because the algorithms minimise the loss function, we are interested in the values of this loss functions at the end of the iterative process. Figure 4.29 shows the comparison between the point reached with the Improved M&Y algorithm and the point reached with the algorithm proposed by Maronna and Yohai (2008) for those 1000 data sets simulated.

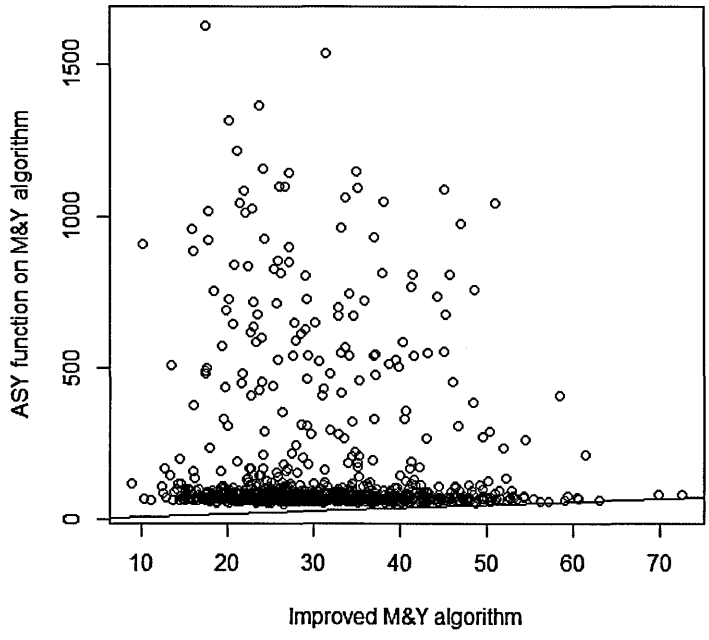


Figure 4.29: Minimal values of the loss function with the Improved algorithm and M&Y algorithm. The line plotted is $y = x$.

The values reached by the loss function with the Improved M&Y algorithm are most of the time smaller than the minimal values reached with the algorithm proposed by Maronna and Yohai (2008). This shows that the point reached by the loss function with our algorithm will be at least as good as the point reached with the method proposed by Maronna and Yohai (2008).

As the semi-Improved M&Y and the Improved M&Y algorithms seem to be very similar we analyse their correspondent minimal values of the loss function.

Figure 4.30 shows that the values reached by the loss function with the Improved M&Y algorithm are most of the time smaller than the minimal values reached with the semi-Improved algorithm.

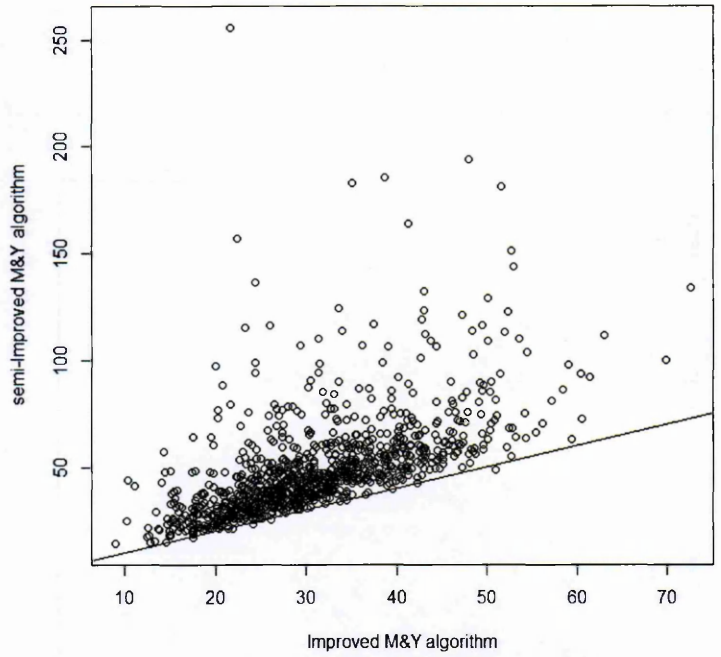


Figure 4.30: Minimal values of the loss function with the Improved and semi-Improved M&Y algorithm. The line plotted is $y = x$.

In conclusion the Improved M&Y algorithm produces good estimates of the vectors **a** and **b** , does not need specific initial values and the value reached by the loss function after the iterations is a good minimal point. This make our proposed algorithm an improvement of the method proposed by Maronna and Yohai (2008) for data sets with asymmetric outliers.

4.6 Aspects of sensitivity of the Improved M&Y algorithm

In this section we analyse the performance of the Improved M&Y algorithm when the data set has more dispersed observations (Subsection 4.6.1) and when the outliers

are symmetric (Subsection 4.6.2). Neither of these cases relates to the orienteering data behaviour, but this analysis will gave an idea of how sensitive the Improved M&Y algorithm is to different possible characteristics of data sets.

4.6.1 Sensitivity to dispersion

In this subsection we analyse the performance of the Improved M&Y algorithm when the data set has more dispersed observations, making the identification of outliers more difficult. To obtain more dispersed observations we change the parameters in the simulation distribution for the vector **b**. So instead of being $b_j \sim$ Uniform(0,0.5) as was for the cases 1 to 4, we use $b_j \sim$ Uniform(5,10). Then based on the simulation method described in Section 4.1, we set the following case:

CASE 5: $a_i \sim$ Uniform(10,20), $b_j \sim$ Uniform(5,10), $e_{i,j} \sim N(0, b_j^2)$, with 20% of outliers

Figure 4.31 shows the dispersion across the columns as we plotted for the other cases, it can be observed that the boxplot is not detecting the 20% of outliers as we wanted. This could be caused by the fact that the values of b_j are larger, so the errors $e_{i,j}$ are bigger (because $e_{i,j} \sim N(0, b_j^2)$). We investigated whether this masking of the outliers has an effect in the performance of the Improved M&Y algorithm. So we compare the estimates obtained with different objective functions in this scenario. We also compare with the results from the algorithm proposed by Maronna and Yohai (2008).

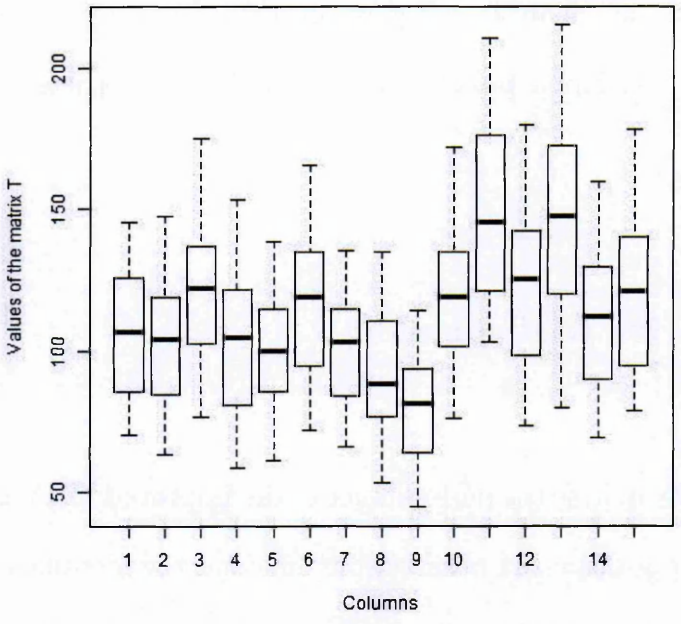


Figure 4.31: Distribution of the simulated data per column, including outliers.

The following table presents the normalized bias and NMSE of the vector estimates obtained with each of the objective functions (ASY, least squares, Huber and bisquare) for 1000 simulated data sets.

		ASY	LS	HUB	BIS	M&Y
CASE 5						
	a NBIAS	0.024	0.032	0.025	0.024	0.017
	a NMSE	0.001	0.001	0.001	0.001	0.000
	b NBIAS	0.000	0.000	0.000	0.000	0.000
	b NMSE	0.000	0.000	0.000	0.000	0.000
	weight bias	-0.155	-0.200	-0.142	-0.096	-0.010
	weight sd	0.004	0.000	0.005	0.006	0.009

ASY: Asymmetric objective function with the Improved M&Y algorithm
LS: Least squares objective function with the Improved M&Yalgorithm
HUB: Huber objective function with the Improved M&Y algorithm
BIS: Bisquare objective function with the Improved M&Y algorithm
M&Y: Bisquare objective function with the algorithm proposed by Maronna and Yohai

Table 4.7: Results of 1000 simulations to compare objective functions performance when the outliers are masked by the observations' dispersion.

Table 4.7 shows that for this case all the functions performed very similarly, there is a small difference between the least square function and the asymmetric, Huber and

bisquare functions. Based on these results we can conclude that there is not enough evidence to suggest that any of the objective functions performs better, and because the values are so close to the least squares it could be said that the algorithm is not detecting the outliers.

We can say that in a case where the data is so dispersed that the outliers are masked, the Improved M&Y algorithm with the asymmetric objective function will perform as well as if we use a symmetric objective function. In this case the method proposed by Maronna and Yohai (2008) (the last column in Table 4.7) has the smallest normalized bias and mean square error and it also has the smallest bias in the weights. So we analysed the weights for the bisquare function in the Improved M&Y algorithm and in the algorithm proposed by Maronna and Yohai (2008) (the last two columns in Table 4.7). We also include for comparison purposes the results for the Asymmetric objective function with the Improved M&Y algorithm.

We analysed the misclassification of outliers and non outliers. So we calculate two separate weight biases, one for those weights whose optimal weight (**W0**) is zero. The other bias correspond to those weights whose optimal weight is one. Table 4.8 presents the results.

Optimal weight		
	0	1
ASY	-0.821	0.011
BIS	-0.727	0.062
M&Y	-0.526	0.119

Table 4.8: Misclassification of outliers on the Improved M&Y and M&Y algorithms. Using 1000 simulated data sets.

The bias of the weights when the optimal weight is equal to zero are large values for the three cases, this suggests that both algorithms have problems detecting all

the outliers, and the algorithm proposed by Maronna and Yohai (2008) seems to be more successful at this task. However the results also show that the Improved M&Y algorithm is more effective with the non outliers, and even more when the objective function is the asymmetric function.

4.6.2 Sensitivity to symmetric outliers

We mentioned in Section 3.2 that the ASY function was developed to deal with asymmetric outliers. We want to study the performance of our asymmetric objective function in cases where the outliers have a symmetric distribution. To do that we have included two more cases in the simulations.

Accordingly, based on method described in Section 4.1, we set $a_i \sim$ Uniform(10,20), $b_j \sim$ Uniform(0,0.5), 20% of outliers and:

CASE 6: $e_{i,j} \sim N(0,1)$, with symmetric outliers.

CASE 7: $e_{i,j} \sim N(0, b_j^2)$, with symmetric outliers.

Table 4.9 shows that for symmetric outliers the asymmetric (ASY) objective function has a small normalized bias, however, it is significantly larger compared with the normalized bias of the other three functions. The normalized mean square error for the ASY objective function is considerably larger than those obtained from the estimates using a Huber, the bisquare and even the least square function. These results suggest that the ASY function will not perform well when the outliers are symmetric. The data set could be box plotted by column before applying the algorithm with an ASY function, to verify the outliers are asymmetric. Also for the particular data sets that we have being studying, is known that the elements of the vector **a** tend to be similar between each other (because they are speeds for a

		ASY	LS	HUB	BIS	M&Y
CASE 6						
	a NBIAS	-0.626	-0.026	0.000	0.000	0.001
	a NMSE	0.408	0.001	0.000	0.000	0.000
	b NBIAS	0.023	-0.026	-0.015	-0.010	-0.015
	b NMSE	0.061	0.007	0.003	0.002	0.004
	weight bias	-0.039	-0.190	0.024	0.030	-0.006
	weight sd	0.018	0.003	0.003	0.003	0.003
CASE 7						
	a NBIAS	-0.664	-0.022	-0.001	0.000	0.000
	a NMSE	0.458	0.001	0.000	0.000	0.000
	b NBIAS	-0.040	-0.004	0.000	0.000	0.000
	b NMSE	0.041	0.000	0.000	0.000	0.000
	weight bias	0.017	-0.190	0.002	0.035	-0.005
	weight sd	0.003	0.003	0.005	0.003	0.002

ASY: Asymmetric objective function with the Improved M&Y algorithm
LS: Least squares objective function with the Improved M&Yalgorithm
HUB: Huber objective function with the Improved M&Y algorithm
BIS: Bisquare objective function with the Improved M&Y algorithm
M&Y: Bisquare objective function with the algorithm proposed by Maronna and Yohai

Table 4.9: Results of 1000 simulations to compare objective functions performance when the outliers are symmetric.

course). So for this simulation cases if the estimates of the vector **a** are negative and/or have a large standard deviation, this might suggest that the outliers are not asymmetric.

The last two columns show the results for the bisquare function with our proposed algorithm and with the algorithm proposed by Maronna and Yohai (2008). Similar values of the normalized bias and mean square error for both algorithms suggest the improvements done to the algorithm might also be valid for symmetric outliers, using a symmetric objective function as the bisquare. But the results show that the weight bias measure is higher for the Improved M&Y algorithm. For this reason an additional analyse of these weights is done by studying the misclassification of outliers and non outliers in the same way we did in Subsection 4.6.1.

The results in Table 4.10 show that in both cases (6 and 7) the Improved M&Y algorithm is better than the algorithm proposed by Maronna and Yohai (2008) in the outlier detection. But the algorithm proposed by Maronna and Yohai (2008) seems to assign weight closer to one to the non outlier observations.

	Optimal weight		
		0	1
CASE 6			
	BIS	-0.028	0.044
	M&Y	-0.104	0.017
CASE 7			
	BIS	-0.004	0.044
	M&Y	-0.044	0.004

Table 4.10: Misclassification of outliers on the Improved M&Y and M&Y algorithms. Using 1000 simulated data sets.

The results in Subsection 4.5.4. suggested that the values reached by the loss function using the Improved M&Y algorithm should be smaller than the values reached with the algorithm proposed by Maronna and Yohai (2008).

	Final value of the loss function		
		mean	standard deviation
CASE 6			
	BIS	1376	137
	M&Y	3284	229
CASE 7			
	BIS	108	32
	M&Y	1631	146

Table 4.11: Analysis of the final value of the loss function on the Improved M&Y and M&Y algorithms. Using 1000 simulated data sets for each case.

The results in Table 4.11 show that for data sets with symmetric outliers the values reached by the loss function using the Improved M&Y algorithm are also smaller than the values reached with the algorithm proposed by Maronna and Yohai (2008), this suggest that the residuals are smaller with the improved algorithm, which means that the estimates are closer to the real values when the Improved M&Y algorithm is used.

However the motivation of this work was to propose a method that performs well in the particular case of asymmetric outliers, and the results in Table 4.5 show that the use of ASY as the objective function produces estimates slightly better than the ones obtained with the bisquare and better than those from the Huber function.

From the analysis in this chapter it is possible to conclude that for data sets with asymmetric outliers, the Improved M&Y algorithm (with ASY, HUB or BIS) produces estimates with smaller normalized bias and mean square error than those given by method proposed by Maronna and Yohai (2008). For the case where the objective function is the asymmetric function, it seems that with the Improved M&Y algorithm, the loss function reaches a good minimum value and the identification of outliers is better.

Chapter 5

Performance measures

As mentioned in Chapter 2 the problem to be solved is to find a procedure that from the matrix of times, estimates the fitness and navigational performance for each of the orienteers running that course. The approach taken in this thesis is through robust lower rank approximation. The algorithm to obtain this approximation of matrices was described in its general form in Chapter 3. Now we will discuss how the lower rank approximation is used in orienteering.

Our main interest is to measure competitor performance in an orienteering event. There are some computer applications that analyse an orienteer's performance after an event, examples of those are Attackpoint and Winsplits (`http://www.attackpoint.org/` and `http://obasen.orientering.se/winsplits/online/en/default.asp`). Winsplits uses a percentage of the best times per leg to estimate a performance index. This index is the quotient between the fastest times per leg and the orienteer time. The median value of the runners' performance indices over the whole course, weighted by the leg length, is considered the runners normal performance for the course. Then if the real time is 25% or 30 seconds over the normal performance for each leg, they will say a mistake has been made. The time lost on that leg will be

the difference between the real time and the estimated normal performance (Troeng, 2008). Attackpoint is a performance and training tool for orienteering athletes, and they have a section dedicated to the analysis of the split times. Attackpoint uses a similar procedure as Winsplits to estimate the expected times and time lost per leg (Attackpoint, 2004).

This way of calculating the runner's performance has the disadvantage that it cannot be computed if the leg length is unknown. Winsplits and Attackpoint use the leg lengths given by the course designers to do their estimations. For this reason, those methods do not consider the fact that the speed will be modified if the distance on the ground differs for each orienteer. That is usually the case as every orienteer can choose between all the multiple routes between check points. Another characteristic of this runner's performance index is that it depends on the orienteers running the course, which suggests that this performance will represent information about the orienteer fitness but it might include features about the course as well. Also the only navigational measure derived from the applications such as Attackpoint and Winsplits is the estimate of time lost.

In this Chapter we intend to present an alternative procedure to estimate an orienteer's performance. We have mentioned that the performance in an orienteering event depends on both the fitness and the navigational skills of the orienteer, so we define our performance measures considering these two aspects. We start in Section 5.1 with an explanation of the estimates obtained from application of the Improved M&Y algorithm mentioned in Chapter 3. From those estimates and other information such as the weights we construct in Section 5.2 two fitness measures and in Section 5.3 three navigational measures. We use data from the Scottish 6

days 2013 event to show how these measures work.

5.1 Output of the lower rank approximation

The robust lower rank approximation described in Chapter 3 was developed to transform the information in the matrix of times into data that provides information about the orienteers performance. The bilinear model provides two vectors **a** and **b**. When these vectors are multiplied, the resulting matrix **H** corresponds to the times orienteers should have taken to complete each leg if they have not got lost in any of the legs (the outliers).

$$\mathbf{T} = \begin{pmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,m} \\ t_{2,1} & t_{2,2} & \dots & t_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n,1} & t_{n,2} & \dots & t_{n,m} \end{pmatrix} \Rightarrow \mathbf{ab^t} \Rightarrow \mathbf{H} = \begin{pmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,m} \\ h_{2,1} & h_{2,2} & \dots & h_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n,1} & h_{n,2} & \dots & h_{n,m} \end{pmatrix}$$

So we have the vector **a** that is related to the speed of the orienteers during the course, the vector **b** that relates to the distance of the legs respectively, and a matrix **H** of estimated times without mistakes.

We will illustrate the outputs of the algorithm with an example. The data used correspond to the times of the orienteers running a long course on the first day of the Scottish 6 days 2013 event. The matrix of original times for the complete data has 143 rows and 14 columns, the following matrix **T** represents only a selection of 6 orienteers out of the 143 that have run the course and is the same that was presented in Table 2.1 in Chapter 2:

$$\mathbf{T} = \begin{pmatrix} 1.18 & 0.78 & 1.80 & 1.67 & 2.78 & 1.20 & 2.65 & 2.38 & 11.33 & 1.00 & 3.38 & 3.72 & 1.07 & 0.43 \\ 1.12 & 0.95 & 1.58 & 1.85 & 2.67 & 1.47 & 3.52 & 2.85 & 10.20 & 0.92 & 3.67 & 4.30 & 1.17 & 0.43 \\ 1.50 & 1.27 & 2.05 & 4.90 & 3.72 & 1.45 & 3.63 & 3.88 & 12.63 & 1.35 & 5.32 & 20.18 & 1.57 & 0.52 \\ 2.82 & 5.83 & 1.82 & 2.52 & 3.27 & 1.88 & 6.33 & 2.78 & 11.00 & 2.67 & 5.43 & \mathbf{25.48} & 1.70 & 0.47 \\ 4.22 & 4.00 & 4.32 & 5.88 & 4.17 & 3.37 & 8.82 & 5.62 & 27.18 & 5.03 & 8.93 & 7.77 & 2.05 & 0.50 \\ 3.32 & 2.78 & 4.98 & 5.58 & \mathbf{14.07} & 3.42 & 9.00 & 5.22 & 28.05 & 7.62 & 12.87 & 18.72 & 2.17 & 0.58 \end{pmatrix}$$

This example corresponds to a course with 14 legs and a total distance of 4775 metres. We can see from visual inspection for example that the $4th$ orienteer did a time of 25.48 minutes on the twelfth leg. Also the last orienteer completed the fifth leg in 14.07 minutes. Looking at the values on the matrix $\mathbf{T}$, these two mentioned times seem to be examples of outliers in this case and hence are marked in bold. The application of the robust lower rank approximation algorithm described in Chapter 3 gave as a result the following two vectors:

$$\mathbf{a} = \begin{pmatrix} 0.006 \\ 0.008 \\ 0.010 \\ 0.010 \\ 0.012 \\ 0.023 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 191 \\ 140 \\ 213 \\ 262 \\ 336 \\ 178 \\ 424 \\ 330 \\ 1174 \\ 174 \\ 499 \\ 664 \\ 147 \\ 43 \end{pmatrix}$$

In Section 3.3 was mentioned that the algorithm will normalize the vector $\mathbf{b}$ with the sum of the values of the vector D. Then here the vector $\mathbf{b}$ is the estimated distances for each leg in metres, and the sum of all the values of the vector $\mathbf{b}$ correspond to the course distance of 4.775 km. As the elements of matrix $\mathbf{T}$ are times then using the physical relationship between speed, distance and time it results that the vector $\mathbf{a}$ represents the inverse of the estimated orienteer's speed, measured in metres per minute. This is orienteer's "pace" (the number of minutes to cover a metre). From

these values it can be seen that the ninth leg was the longest with 1174 metres followed by the twelfth with an estimated distance of 664 metres. The $1st$ orienteer is estimated to be the fastest, the $6th$ the slowest and the $3rd$ and $4th$ have similar speeds.

Then the estimated times given by the multiplication of those two vectors are:

$$\mathbf{H} = \begin{pmatrix} 1.15 & 0.84 & 1.28 & 1.57 & \textit{2.02} & 1.07 & 2.54 & 1.98 & \textit{7.04} & 1.04 & 2.99 & 3.98 & 0.88 & \textit{0.26} \\ 1.53 & 1.12 & 1.70 & 2.10 & 2.69 & 1.42 & 3.39 & 2.64 & 9.39 & 1.39 & 3.99 & 5.31 & 1.18 & 0.34 \\ 1.91 & 1.40 & 2.13 & \textit{2.62} & 3.36 & 1.78 & 4.24 & 3.30 & 11.74 & 1.74 & 4.99 & \textit{6.64} & 1.47 & 0.43 \\ \textit{1.91} & \textit{1.40} & 2.13 & 2.62 & 3.36 & 1.78 & \textit{4.24} & 3.30 & 11.74 & \textit{1.74} & 4.99 & \textbf{\textit{6.64}} & 1.47 & 0.43 \\ \textit{2.29} & \textit{1.68} & \textit{2.56} & \textit{3.14} & 4.03 & \textit{2.14} & \textit{5.09} & \textit{3.96} & \textit{14.09} & \textit{2.09} & \textit{5.99} & 7.97 & 1.76 & 0.52 \\ 4.39 & 3.22 & 4.90 & 6.03 & \textbf{\textit{7.73}} & 4.09 & 9.75 & 7.59 & 27.00 & \textit{4.00} & \textit{11.48} & \textit{15.27} & 3.38 & 0.99 \end{pmatrix}$$

These are the times we expected each orienteer to have taken to complete each leg, based on him/her running at constant speed along the course and not making any navigational mistake. Marked with *italics* are the times that the algorithm identifies as outliers. So we have that the expected time of the fourth orienteer on the twelfth leg is 6.64 minutes instead of the 25.48 minutes he/she actually did. For the last orienteer the expected time on the fifth leg was 7.73, that compared with the original times in the matrix $\mathbf{T}$ is almost half the time. These large differences between the original times and the expected suggest the orienteers got lost in those legs.

The matrices $\mathbf{T}$ and $\mathbf{H}$ and the vectors $\mathbf{a}$ and $\mathbf{b}$ are used to generate an evaluation of the fitness and the navigational performance for each orienteer on the course. In the following subsections the construction of two fitness and three navigational measures are explained.

5.2 Fitness measures

The reciprocals of the elements of the vector **a** will be the average speed of the orienteers over the course. This is the expected speed with a perfect navigation. So if an orienteer runs a course of ten legs and got lost in the third and seventh leg, his/her speed will be calculated considering the expected time without making a mistake on legs three and seven.

The average course speed of the ith orienteer is our first fitness measure and is defined as:

$$speed_i = \frac{1}{a_i} \tag{5.1}$$

where a_i is the i-th element of the vector **a**.

It is important to mention that this speed is related to how runnable the terrain where the event took place was. So courses in areas with open land will produce faster average speeds than a course in an area with forest that is difficult to run through. So the difference in average speed between orienteering events of the same length is non-negligible unlike other types of running races.

In order to be able to compare the performance of the same orienteer over different events we need to eliminate the relationship of the speed with the course. To do this we define a ranked speed. This is a comparison of each orienteer's average speed with the mean speed of the group that only includes the fastest 10% of the orienteers running that course. The 10% was chosen to have a ranking that does not depend only on one runner (the fastest). The fastest orienteer changes from course to course so using a group instead will make the ranked average speed a more stable measure,

not dependent so much on the orienteers participating on the course. If the group only includes the best orienteers on the course, the ranked average speed will be in terms of the best possible times for the course, measuring how far the orienteer is from the winning times. A greater percentage like 20% or 50% will increase the variance of the speed in the group, which might cause that the dependence of the speed on the course is not eliminated as desired for the ranked average speed. For this reason we decide to have a small group (10%). The ratio is multiplied by 100 to have it in percentage terms.

So the ranked average speed for the i-th orienteer is defined as:

$$fit_i = \frac{speed_i}{mean_{best10\%}(speed_i)} \times 100 \tag{5.2}$$

Continuing with the example of the selected 6 orienteers from a course on the first day of the Scottish 6 days 2013 event, the mean speed of the group that only includes the 10% of the fastest orienteers in this course is 128 metres per minute. Their correspondent fitness measures are:

$$speed = \begin{pmatrix} 167 \\ 125 \\ 100 \\ 100 \\ 83 \\ 43 \end{pmatrix} \qquad fit = \begin{pmatrix} 130 \\ 98 \\ 78 \\ 78 \\ 65 \\ 34 \end{pmatrix}$$

The first orienteer ran the course at an average speed of 167 metres per minute in contrast, the fifth orienteer had an average speed of 83 metres per minute. The ranked average speed of 130 for the first orienteer means he/she is running faster than the mean of group with the 10% of the fastest orienteers. The sixth orienteer

has a ranked average speed of 34, this means that this orienteer's speed is only 34% of the speed of best 10% orienteers.

5.2.1 Comparison of the average speed measure

To investigate whether these fitness performance measures are appropriate we will compare our average course speed with running speeds found in the literature. The relation between age and running speed has been studied for different running events. One of these studies was done by Bird et al. (2001), who showed that the running speeds for a cross-country event for the fastest men and women between the ages of 45 to 65 years decrease linearly by an average between 7% and 9% per decade of age. Korhonen et al. (2003) concluded that the sprint performance decline on average between 5% to 6% per decade in males and between 5% to 7% per decade in females. In a particular case for orienteering events, Bird et al. (2001) showed that the speed for men and women between 50 and 70 years decreased by 17% and 21% per decade.

The relation between running speed and gender has also been studied by Gjerset et al. (1997). Their research shows that women's speed is 22% slower than that of men in a cross-country race, compared with a 32% difference in orienteering terrain.

We want to analyse whether our average speed measure behaves similarly to what has been reported in the papers mentioned. We use the data from 14 different courses on the first day of the Scottish 6 days 2013 event to analyse the average course speed for each age group and both genders. The courses selected were the long courses for the 35 to 80 age groups for men and women. Bird et al. (2001) used the fastest three orienteers in each group for their analysis. So for this analysis we used as well the three fastest orienteers at each age group. We used the average

speed measure to identify the fastest orienteers. The speed estimated for the 40 years old group is used as the starting point to calculate the speed decline for cross country and orienteering. These estimates are based on the declines reported in the literature. The results are plotted in the following Figure 5.1.

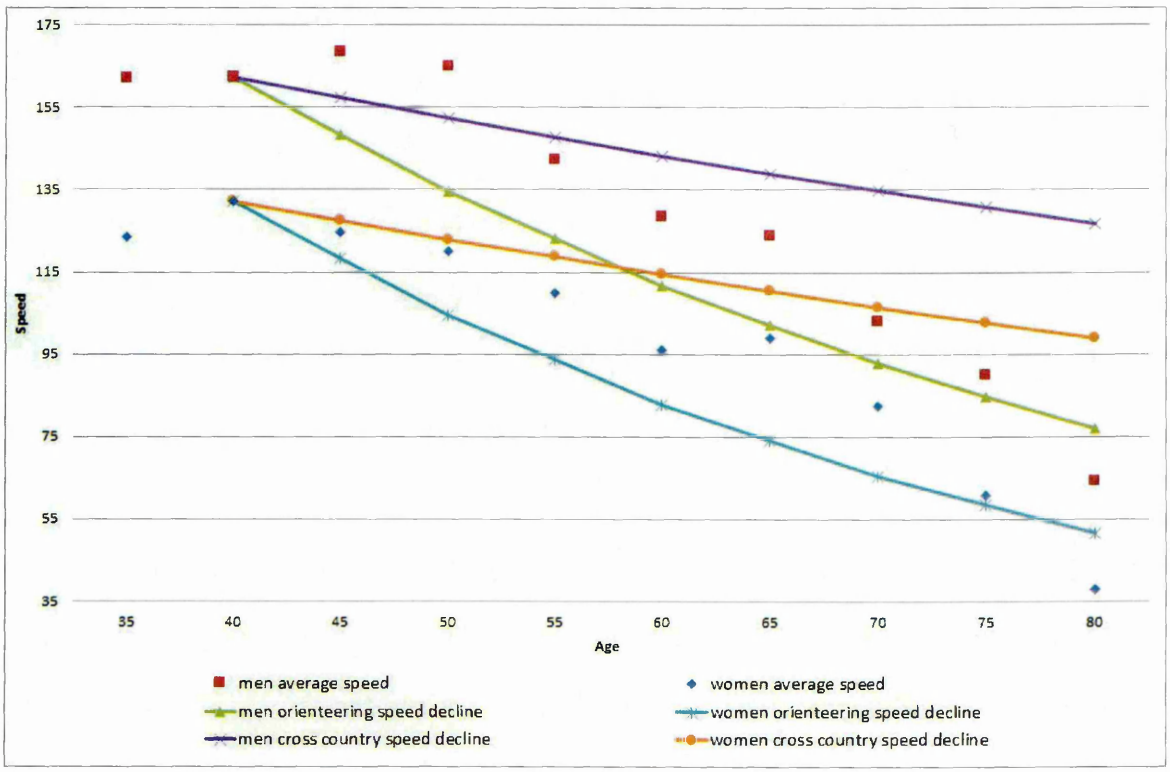


Figure 5.1: Comparison of average course speed by age and gender against the running speed for the cross country and orienteering events based on the decline reported in the literature.

Figure 5.1 shows that the average speed of the competitors running the first day of the Scottish 6 days 2013 event declines from the age groups of 45 and 40 years old onwards for men and women respectively. In particular the decline observed from 40 years old to 80 years old has the same tendency as the speed decline by age found in the literature for cross country and orienteering events. The differences of speeds between genders mentioned by Gjerset et al. (1997) is also present in this data and shown in this figure. The fact that our average speed follows the tendency described in the literature suggest that this fitness measure is appropriately estimating an orienteer's speed.

5.2.2 Analysis of the ranked average speed measure

As mentioned we propose the use of the ranked average speed as a fitness measure that allows us to make a comparison of the competitors fitness across different events. To verify that this measure allows the comparison, we use the data from one specific colour coded course (named green course) over 6 days of the same Scottish event already mentioned. The green course has a high technical difficulty, a maximum length of 5 km, and is open for orienteers of any age. Figure 5.2 presents a matrix scatter plot of the ranked average speed for the 36 orienteers that participated in all 6 days of competition always on a green course. The measure was calculated for each day based on all the orienteers running the green course.

We do not expect an orienteer's fitness to change between days of the event, unless an injury had occurred. Figure 5.2 shows the relation between the ranked average speed on the six different days. On the lower panels we observe the scatter plots for each pair of the days, the diagonal lines correspond to the identity function. The upper panels present the correlation coefficient for each case. The plots suggest that this fitness measure is very similar in all the days. This fact supports the hypothesis that the ranked average speed is measuring the orienteers' fitness, and that is a measure that can be use to compare fitness across courses.

The ranked average speed has also to be consistent with the original times. We expect an orienteer with a large original time to have a low ranked average speed and vice versa for short original times to complete the course. To corroborate that relationship between this fitness measure and the original times, in Figure 5.3 we plotted the ranked average speed against the original times of the same 36 orienteers for each of the 6 days.

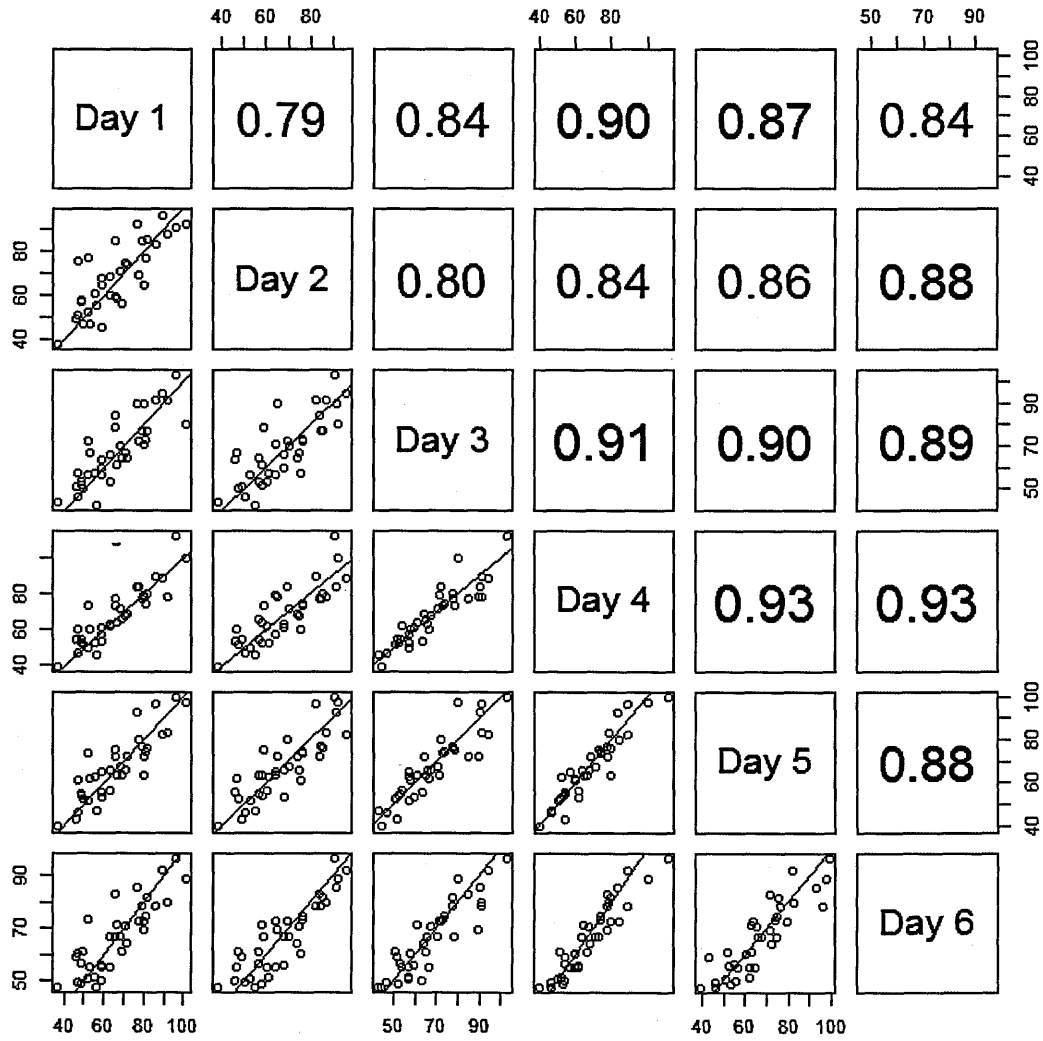
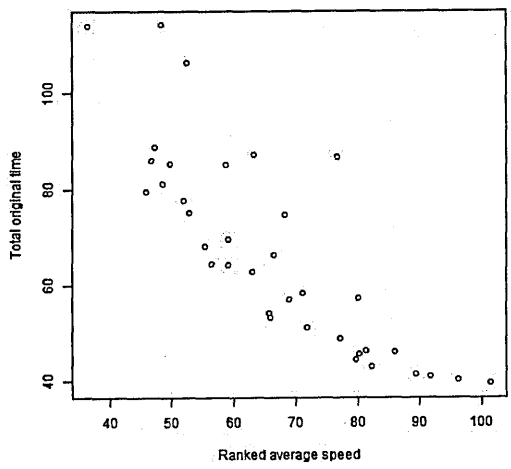
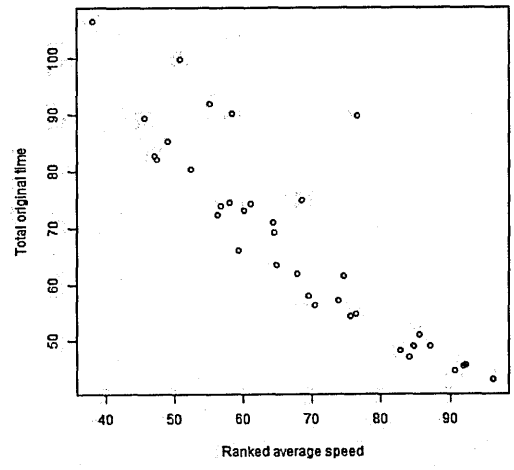


Figure 5.2: Comparison of ranked average speed for orienteers running a green course on each of the 6 days of the Scottish 2013.

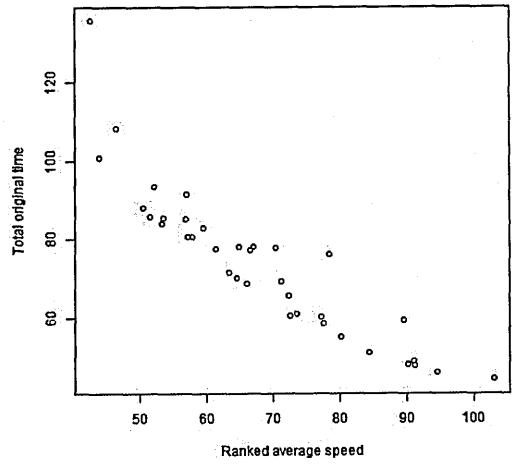
The graphs in Figure 5.3 show that the original times and the ranked average speed have a negative correlation as was expected. Most of the observations lie on a visible curve. The observations far from the main curve correspond to the orienteers that on that day made mistakes that affected their original times. So we can conclude that the ranked average speed is measuring an orienteer's fitness as desired.



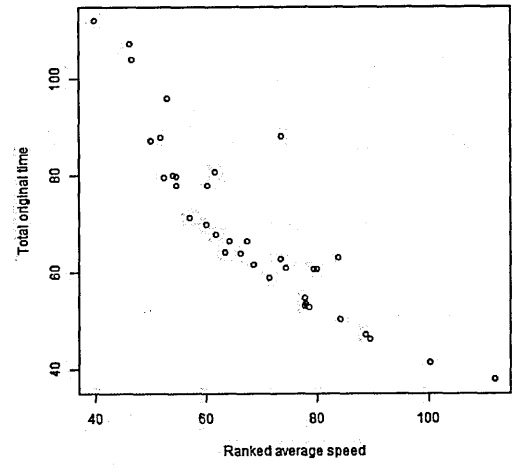
(a) Day 1



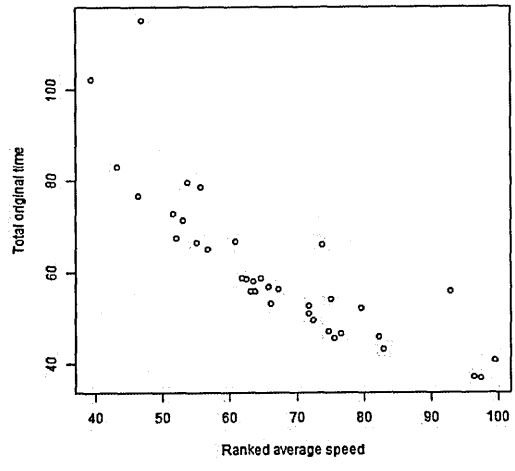
(b) Day 2



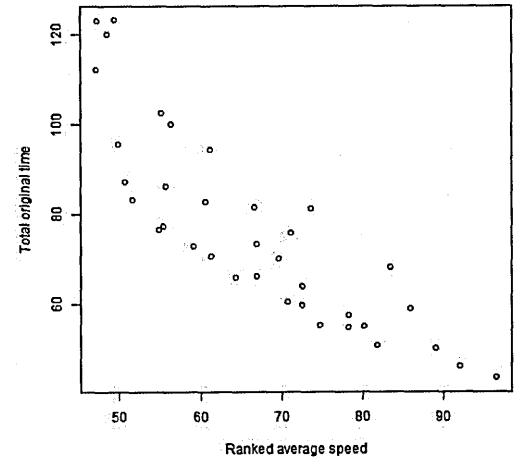
(c) Day 3



(d) Day 4



(e) Day 5



(f) Day 6

Figure 5.3: Comparison of ranked average speed for orienteers running a green course on each of the 6 days of the Scottish 2013.

By construction the ranked average speed depends on the top 10% of each particular course, but it could be possible to choose a different top 10% that does not depend on the course. For example we can use courses run in the same area, the same day and that by design have similar technical difficulty (such as the green and blue course in one event). In such case we will expect that the speeds of the competitors in the different courses should be comparable, and instead of choosing the top 10% of each colour we calculate the ranked average speed with the fastest 10% over the two colour coded courses. Then this ranked speed will be dependent on the area and not on the courses.

As we will show in Chapter 6, the data suggest that the assumption of two courses having similar physical and navigational level is a strong assumption. For this reason we used the top 10% of the course, and we have shown that the ranked average speed defined this way fit the purpose of producing a fitness measure that is comparable between courses.

5.3 Navigation

To measure the navigational skill we use three different variables, the speed consistency, the number of legs with mistakes and the time lost in those mistakes. There is a correlation between these variables as all of them depend on the extra time taken by the orienteer in certain legs, but they complement each other because they are measuring different aspects of the navigational skill.

The robust lower rank approximation algorithm through the asymmetric objective function assigns weights $w_{i,j}$ to each time in the matrix. As these weights are directly related to the time the ith orienteer took to complete the jth leg ($t_{i,j}$), the

weights are also related with outliers. So it is possible to use the weights to identify navigational mistakes, and estimate the number of legs with mistakes as well as the time lost in those mistakes.

The asymmetric objective function will assign weights between zero and one to the times, giving a zero weight to large outliers. So we define that if a weight value is smaller than 0.75 then we will consider that the orienteer took longer than expected for that leg. This means that the time $t_{i,j}$ is detected as a time where a navigational mistake has been made.

In Subsection 4.4 was shown that for the simulated case which is similar to the orienteering data (case 4), the use of the 0.75 value in the method appears to detect all the outliers. In this case we will be selecting all the legs where a mistake has been made. The cutting value (c-value) for the asymmetric function discussed in Chapter 3 needs also to be adjusted. For the orienteering data the value of $c = 4$, seems to work well, meaning that it detects all the obvious outliers and the percentage of outliers in a course is on average 20%. So $c = 4$ is the value used for this analysis.

Speed consistency

The speed consistency measures how much the speed varies between legs. We assume that this variation is caused by navigational factors, because the difficulty of completing each leg will affect directly in the time to complete it, so the speed will be different. So orienteers with very consistent speeds along the legs, will have smaller values on this speed consistency measure. Then low values of speed consistency means the orienteer has found the course easy to navigate suggesting he/she has good navigational skills. This measure is estimated with the variation between

the real times **T** and the expected times **H**. For each orienteer and each leg, we estimate how different the real time is from the expected. Then, by adding the absolute value of the variations in all the legs and dividing by the number of legs, we obtain an estimated variation over all the course. Then the speed consistency of the ith orienteer will be:

$$cons_i = \frac{\sum_{j=1}^{m} |\frac{t_{i,j}}{h_{i,j}} - 1|}{m}. \tag{5.3}$$

So the higher this value is, the less consistent the orienteer's speed was.

Proportion of legs with mistakes

We compute the proportion of legs per orienteer that were detected as having a navigational mistake. This is done with the matrix of weights used in the robust bilinear fitting as mentioned in previous paragraphs.

$$pro\_leg_i = 100 \times \frac{\sum_{j=1}^{m} I_{i,j}}{m} \tag{5.4}$$

where

$$I_{i,j} = \begin{cases} 0 & \text{if } w_{i,j} \geqslant 0.75 \\ 1 & \text{if } w_{i,j} < 0.75 \end{cases}. \tag{5.5}$$

Time lost in mistakes

The time lost in mistakes measures the navigational ability in terms of the amount of time spend on mistakes. The difference between the real time and the estimated time from matrix **H**, gives an estimate of the minutes lost in each leg. If only the legs where mistakes are detected by the algorithm are considered then these differences will be positive. So the estimated minutes lost can be calculated with the equation:

$$min\_lost = \sum_{j=withmistake} t_{i,j} - h_{i,j} \tag{5.6}$$

We are interested in measuring the time lost relative to the length of the leg, because in terms of the navigational performance, 1 minute lost in a 3 minutes leg will not be the same as 1 minute lost in a 10 minutes leg. So the minutes lost in the legs identified as mistakes are weighted in such way that minutes lost in longer legs are down weighted. The weights definition is based on the ratio between each leg distance and the total distance of the course. To obtain a down weight on large distances we use $(1 - ratio)$, so the weights will go from zero to one. We use the estimated leg distances given by the vector **b** of the bilinear model. The time lost by the ith orienteer will be:

$$tlost_i = \sum_{j=withmistake} (t_{i,j} - h_{i,j}) \times (1 - \frac{b_j}{D}) \tag{5.7}$$

where D is the total course distance. Because the weights in the time lost measure take values between zero and one, the value of the time lost in mistakes will always be smaller than the estimated minutes lost.

Following the example of the selected 6 orienteers from a course on the first day of the Scottish 6 days 2013 event their corresponding navigational measures are:

$$cons = \begin{pmatrix} 0.217 \\ 0.123 \\ 0.322 \\ 0.595 \\ 0.614 \\ 0.279 \end{pmatrix} \quad pro\_leg = \begin{pmatrix} 21\% \\ 0\% \\ 14\% \\ 36\% \\ 71\% \\ 29\% \end{pmatrix} \quad tlost = \begin{pmatrix} 4.11 \\ 0.00 \\ 13.81 \\ 24.19 \\ 29.85 \\ 13.59 \end{pmatrix}$$

We have that the second orienteer has a speed consistency of 0.123, this means that he/she was running all the legs at a more consistent speed than the other five runners and he/she did not got lost in any leg as the $pro\_leg = 0\%$ means that the proportion of legs where a mistake was made is zero, similarly for the time lost ($tlost = 0$). The fourth orienteer had a average speed of 100 metres per minute however the consistency of this speed was 0.595, which means that the speed varied from leg to leg, the value of the *pro_leg* means that he/she got lost in 36% of the legs (5 legs) and the value of the *tlost* means that in those 5 legs he/she lost more than 24.19 minutes.

As mentioned in Chapter 4 the algorithm is sensitive to large variation between legs. An example of this is the fifth orienteer, who despite having a speed consistency similar to the fourth orienteer, the estimated proportion of legs where a mistake was made is 71%. However most of the detected mistakes are very small, between 1.5 and 3 minutes in each leg compared with the fourth orienteer which mistakes in legs two and twelve added 4.43 and 18.84 minutes to his/her time. This means that part of the speed variability across the legs is due to the time lost in errors and it can be detected by analysing the three navigational measures together as each one is providing different information about an orienteer's navigation performance.

5.3.1 Analysis of the navigational measures

To analyse the proposed navigational measures we use the data from the first day of the green course of the Scottish 6 days 2013 event, the same data used in the

comparison of the ranked average speed measure in Subsection 5.2.2. Figure 5.4 shows the three navigational measures on the same plot. The graph has the speed consistency and time lost in the axes and the different colours of the points in the scatter plot represent the percentage of legs detected with outliers.

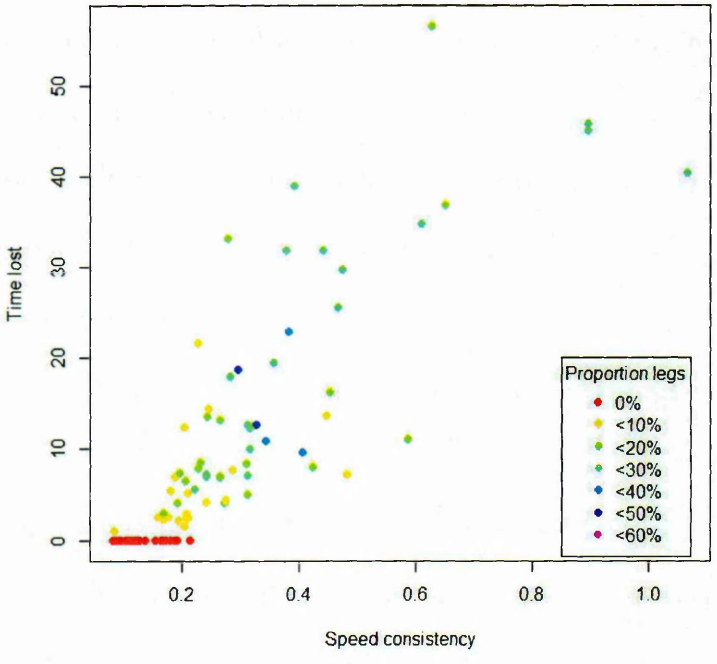


Figure 5.4: Navigation performance measures, for orienteers running a green course on the first day of the Scottish 6 days 2013 event.

Figure 5.4 shows the direct relationship between speed consistency and time lost. Also as expected there is a correlation between proportion of legs with mistakes and time lost. The combination of a high proportion of legs with outliers and low speed consistency usually implies little time lost. In the plot this is true for the two darker blue points in the middle of the scatter plot. This suggests that most of these cases are orienteers with many legs detected with outliers but with very small amounts of time lost. On the other hand high values of time lost with small proportion of legs suggest the orienteer lost a considerable amount of time in few legs, clear navigational mistakes (yellow and green points). Figure 5.4 shows that the three

navigational measures have correlation between each other, however it seems that they measure different aspects of an orienteer's navigational performance. For this reason we propose that the navigational performance analysis is done as a conjunction of the three measures.

In the following Figure 5.5 we use as reference the actual times to analyse the speed consistency and time lost. The scatter plot has the original time and speed consistency in the axes and the time lost is represented with the different colours for the points in the plot.

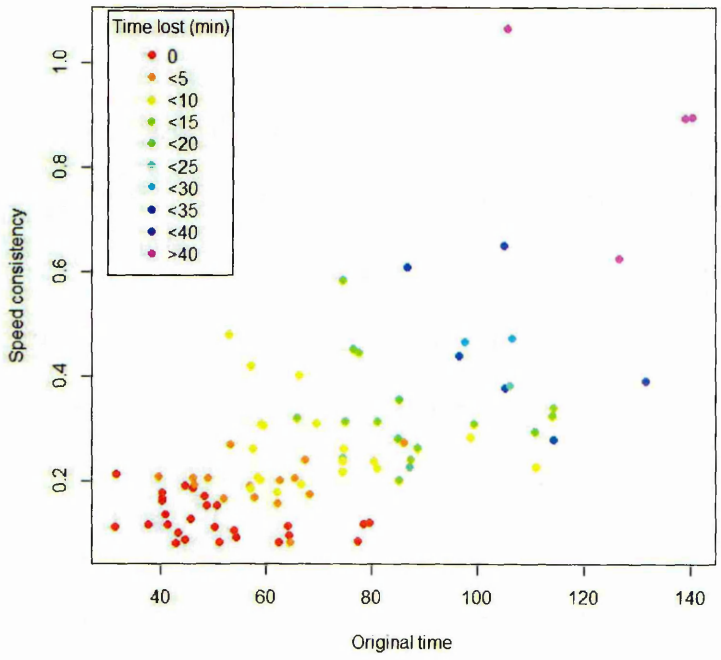


Figure 5.5: Speed consistency and time lost against original times, for orienteers running a green course on the first day of the Scottish 6 days 2013 event.

Remember that the actual times include the times spent on mistakes, so we expect a correlation between high original times and time lost. Figure 5.4 showed there is a relation between speed consistency and time lost. However in Figure 5.5 the pattern of time lost increasing as the speed consistency increases is clearer.

Small actual times are related to low values of the speed consistency and low values of time lost (red points). As the actual time increases, the spread of the speed consistency also increases. This is due to the fact that at this point the model differentiates between two types of orienteers: 1) Slow runners with good navigational skills who will have high actual time and low speed consistency (as their speed will be steady along the course), and 2) average or fast runners with not very good navigational skills that will cause their speeds to vary between legs making the speed consistency measure increase.

From Figures 5.4 and 5.5 is possible to conclude that the navigational measures are correctly detecting important features of an orienteers' navigational skills.

Now we would like to analyse how the navigational measures behave across events. The results of the 36 orienteers that always ran a green course over each of the 6 days Scottish event are the data used for this analysis. The three navigational measures were calculated for each day. To compare the results across the different days we obtain a matrix plot and calculate the correlation matrix for the speed consistency and for the time lost measures.

In Figure 5.6 we present the plots of the speed consistency measure of each day against the other 5 days for the Scottish event. The diagonal lines correspond to the identity function $y = x$. The points in the scatter plots are spread, these suggest that there is no significant correlation between the speed consistency estimated for an orienteer in the course run on one day and the speed consistency obtained by the same orienteer in the courses run the other days. A similar result was shown on the matrix plot for the time lost measure, suggesting also that the time lost in one

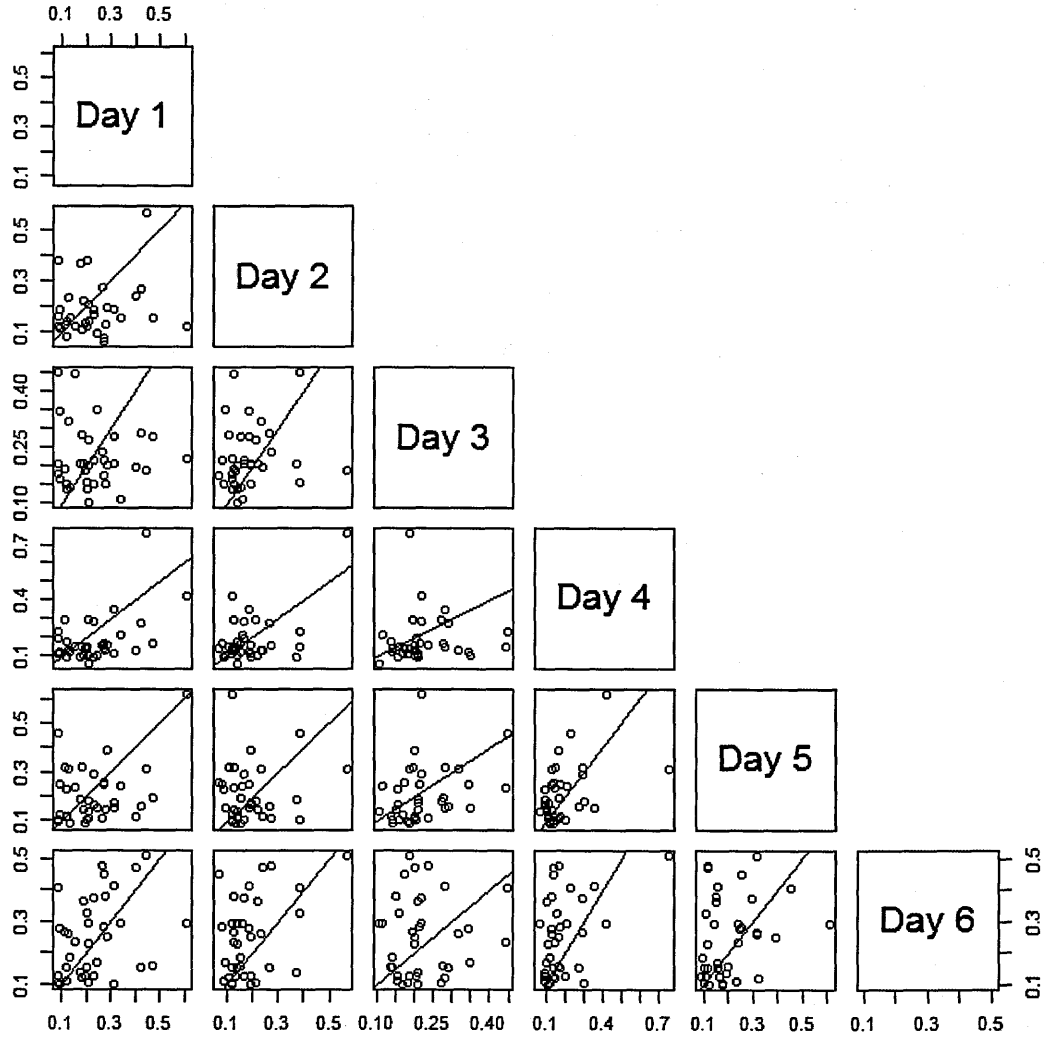


Figure 5.6: Comparison of speed consistency for orienteers running a green course each of the 6 days of the Scottish 2013.

course is independent of the time lost in a different course.

Table 5.1 presents the correlation coefficients for these navigational measures. For the speed consistency the coefficients vary from -0.07 to 0.52. This affirms what was seen in the previous plots that there is not a clear correlation between this navigational measure over different courses. The coefficients suggest the possible correlation on the consistency of speed between the courses on day 4 and the courses on the first and second day.

	Speed consistency					
Day	**1**	**2**	**3**	**4**	**5**	**6**
1	1.00					
2	0.17	1.00				
3	-0.07	0.15	1.00			
4	0.49	0.52	0.08	1.00		
5	0.32	0.11	0.35	0.45	1.00	
6	0.34	0.40	0.07	0.41	0.25	1.00

	Time lost					
Day	**1**	**2**	**3**	**4**	**5**	**6**
1	1.00					
2	0.08	1.00				
3	0.02	0.29	1.00			
4	0.27	0.64	0.35	1.00		
5	0.15	0.39	0.53	0.60	1.00	
6	0.22	0.49	0.40	0.57	0.47	1.00

Table 5.1: Correlation coefficients for speed consistency and time lost across the 6 days of the Scottish event.

The correlation coefficients for the time lost measure shown in Table 5.1 also suggest that this measure represents a characteristic of the course that does not depend on the orienteer. The values of the coefficients suggest a possible correlation between the time lost in day 4 and the time lost on days 2, 5 and 6. Also between the third and fifth day.

This analysis suggests that the navigational measures proposed do not give a consistent measure of the orienteers' technical ability. The measures appear to be highly related to an orienteer's performance in that particular course. This might be because of the factor of making mistakes involved in the measure. Furthermore getting or not getting lost in a leg may not have a clear relationship with an orienteer's innate technical ability. However we manage to obtain information about an orienteer's navigational skills from those mistakes. This suggest that the navigational measures are more variable than the speed over different courses. For this reason the use of a ranked measure as we did for the speed will not provide a measure that is consistent across events. However as shown before focusing on a single course, the three navigational measures provide information that can be used to measure the navigational skills of the orienteers.

An example of how the proposed navigational measures could be used to analyse performance over different events is presented in Figure 5.7. This figure presents the analysis of the speed consistency and time lost measures for 6 orienteers (each orienteer is represented by a colour) on the six days (each point represents a day as labelled) of the Scottish 2013 event. The six orienteers were selected randomly.

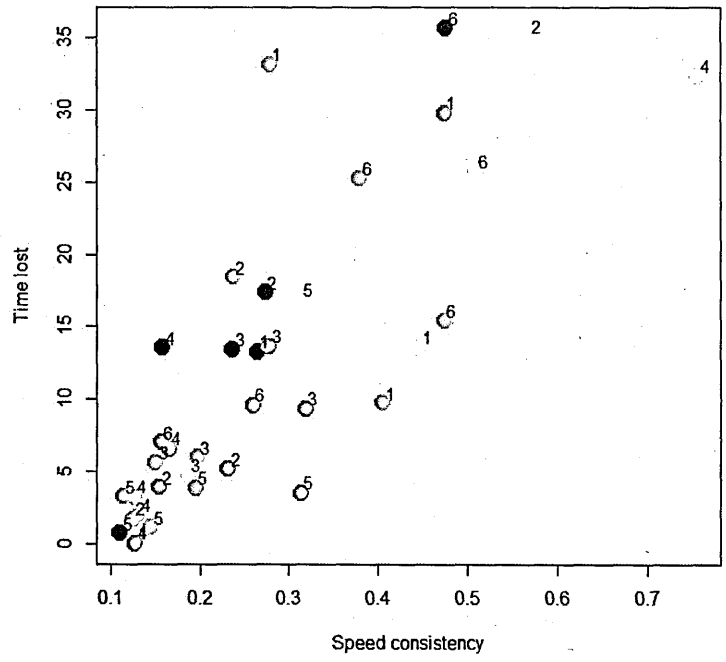


Figure 5.7: Comparison of speed consistency and time lost for 6 orienteers running a green course each of the 6 days of the Scottish 2013.

Figure 5.7 shows that for four orienteers (yellow, green, light blue and blue) the course on day 6 was harder navigationally, which might suggest that this course was different to the courses on the other five days. We also observed that some orienteers (except the yellow and light blue orienteers) have speed consistency and time lost measures fairly close to each other in four or five of the days. Choosing as an example the orienteer marked in pink colour, the results suggest that this orienteer performed very well on the second day, good on the fourth, fifth and sixth days, but on day one made a large mistake that cost him/her a time lost of at least 30 minutes

and the course on day three was more navigationally challenging for him/her. This might suggest that his/her navigational skills are good for the green course. To complete the example, this orienteer's ranked average speeds were between 48 and 57 for the 6 days. This suggest that because his/her navigation level is good for the green course, this orienteer might be able to improve his/her performance in green courses by only improving his/her fitness.

We have constructed two fitness and three navigational measures. The average speed and the ranked average speed measure the fitness level of the orienteer. The average speed also includes information about how easy or difficult to run the course was. The ranked average speed is a measure that allows the comparison of the orienteer fitness across different events. The analysis of the navigational measures; speed consistency, proportion of legs with mistakes and time lost in mistakes together provides a clear view of an orienteer's navigational skills. However, the lack of consistency of the orienteers making mistakes over different courses cause that these navigational measures are only valid for their correspondent course. Still, the combined analysis of these five measures gives a more complete approach to the complex problem of modelling an orienteer's performance.

Chapter 6

Difficulty scale for colour coded events

The courses in an orienteering event are usually organized by class (gender and age) or by colour. For colour coded events in the UK the British Orienteering Federation (BOF) has set the guidelines that event organizers should follow in the design of their courses to ensure the consistency of course standards. These guidelines are based on the length, the technical difficulty level and the elite winning times of the courses. The technical difficulty level is based on the skills needed to successfully complete a course. The levels go from 1 to 5, with 5 being courses requiring the highest technical skills (BOF, 2014).

In the case of the colour coded events, the elite winning times are calculated with a course length ratio based on the black course (the most difficult course). So those times will correspond to the expected times for the top standard elite competitors. In a similar way, age class events use a speed ratio to estimate the course's length, the ratio is based on the speed of the class M21 (men of 21 to 34 years old) (Appendix B to Rules of Orienteering).

The following table presents the suggested values for the more common colour courses on long distance events:

Colour	Distance	Technical Difficulty	Elite winning time
WHITE	1.0 to 1.9 km	1	9 minutes
YELLOW	2.0 to 2.9 km	2	15 minutes
ORANGE	2.5 to 3.5 km	3	17 minutes
LIGHT GREEN	3.0 to 4.0 km	4	20 minutes
GREEN	3.5 to 5.0 km	5	26 minutes
BLUE	5.5 to 7.5 km	5	38 minutes
BROWN	8.5 to 12.0 km	5	57 minutes
BLACK	10.0 to 14.0 km	5	67 minutes

Table 6.1: BOF guidelines for long distance colour coded events

In Table 6.1 we can see that the green, blue, brown and black colours only differ in length as all of them have the highest possible technical difficulty level (5). The distance specified in this table corresponds to a corrected distance, this correction is based in an addition of 0.1 km for every 10 metres of climb along the course.

As mentioned in Chapter 1 the data used for this work are the courses information published on the BOF results web page, and correspond to the matrix of times, length of the course, number of controls and climbing metres. However, according to the guidelines course planners will use the length, the number of controls, the weather and the terrain to obtain the desired difficulty level and estimated elite winning time for the black course. This suggests that using the length, number of controls and climbing metres to identify similarities between same colour courses might not be possible. Nevertheless, the guidelines suggest the elite winning times and technical difficulty should be very similar between same colour courses. The use of elite winning times directly to analyse the course similarities is usually not possible as on some courses, elite orienteers are unlikely to enter. For example an event with green, blue and brown courses, the elite orienteers most probably decide

to run the brown course as it will be more physically demanding, and elite orienteers running a green or blue course might have chosen to do so because of illness or injury that is affecting their performance.

We are interested in studying the consistency of the courses in different events. We compared green, blue and brown courses across different events in the UK and found evidence suggesting differences in the technical difficulties. The results of this comparison are presented in this chapter. First we analyse the original times in each event, and this is followed by studying the courses' technical difficulty based on the performance measures developed in Chapter 5. The chapter finishes with the proposal of a methodology to measure the course technical difficulty.

The analysis was done with data downloaded from the British Orienteering web page. These data correspond to the published results for the green, blue and brown courses of 100 events that took place in the UK between January 2013 and May 2014. The data was gathered in a way such that the course labels in any of the colours corresponds to courses from the same event. This means that the first green course was run the same day and in the same area as the first blue and the brown courses.

6.1 Analysis of the original times

Studying the original times for all the courses will provide a rough idea of the course technical difficulty. This will also let us compare the winning times between courses. Based on the assumption that course planners take into consideration the BOF guidelines regarding the elite winning times, we will expect the winning times for each course (if an elite orienteer was running it) to be close to the times estab-

lished in Table 6.1.

Figures 6.1, 6.2 and 6.3 present the distribution of the original times for each course. The data in each boxplot represent the orienteers running that particular course. The elite winning time as given in Table 6.1 is the horizontal line in each figure.

Figure 6.1 shows that most of the winning times for the green courses are above the recommended 26 minutes for elite orienteers. The mean time per course goes from 42 to 94 minutes, and the overall mean time is 68 minutes.

Figure 6.2 shows that only 8 of the blue courses have winning close to or below the recommended 38 minutes. The mean time per course goes from 50 to 115 minutes, and the overall mean time is 76 minutes.

The times for the brown courses are shown in Figure 6.3. The mean time to complete a brown course is 84 minutes. However across courses the mean time goes from 56 to 125 minutes. Around half of the courses had winning time below or very close to the 57 minutes suggested by the BOF guidelines.

The difference between the winning times and the elite winning times could be caused by the fact that the elite winning times on Table 6.1 are calculated with ratios from the elite winning time for the black course (67 minutes), and not all the events offer a Black course which suggest that planners have to estimate this time. Also events with a larger winning time on the black course will cause larger winning times on all the other courses. Another reason for this difference is the absence of elite runners competing in the analysed courses and this is a common

case for shorter courses. We also know that the definition of elite orienteers depends on orienteers running elite courses, then the identification if whether the winner on a non elite course was an elite runner is complex. These results suggest that using the elite winning times as guide to assess the technical difficulty of the course might not be workable path.

To complete the analysis of the original times we calculated the correlation coefficients between the times and the course characteristics (length, number of controls and climbing metres). The results for the three colours are presented in the following table.

Correlation coefficient	**Times**		
	Green	Blue	Brown
Number of controls	0.081	0.115	0.064
Distance	0.024	0.086	0.183
Climbing metres	0.119	0.147	0.114

Table 6.2: Correlation between original times and course characteristics.

Low values of the coefficients suggest that there is no correlation between the times the orienteers take to complete the course and the distance, number of controls or climbing metres on a course. This is an expected result as it was mentioned that these variables are used by the course designer to determine the technical difficulty of the courses.

The results suggest that the length, number of controls and climbing metres of a course do not provide information about the course technical difficulty level, so in order to assess the difficulty of the courses other variables will have to be used.

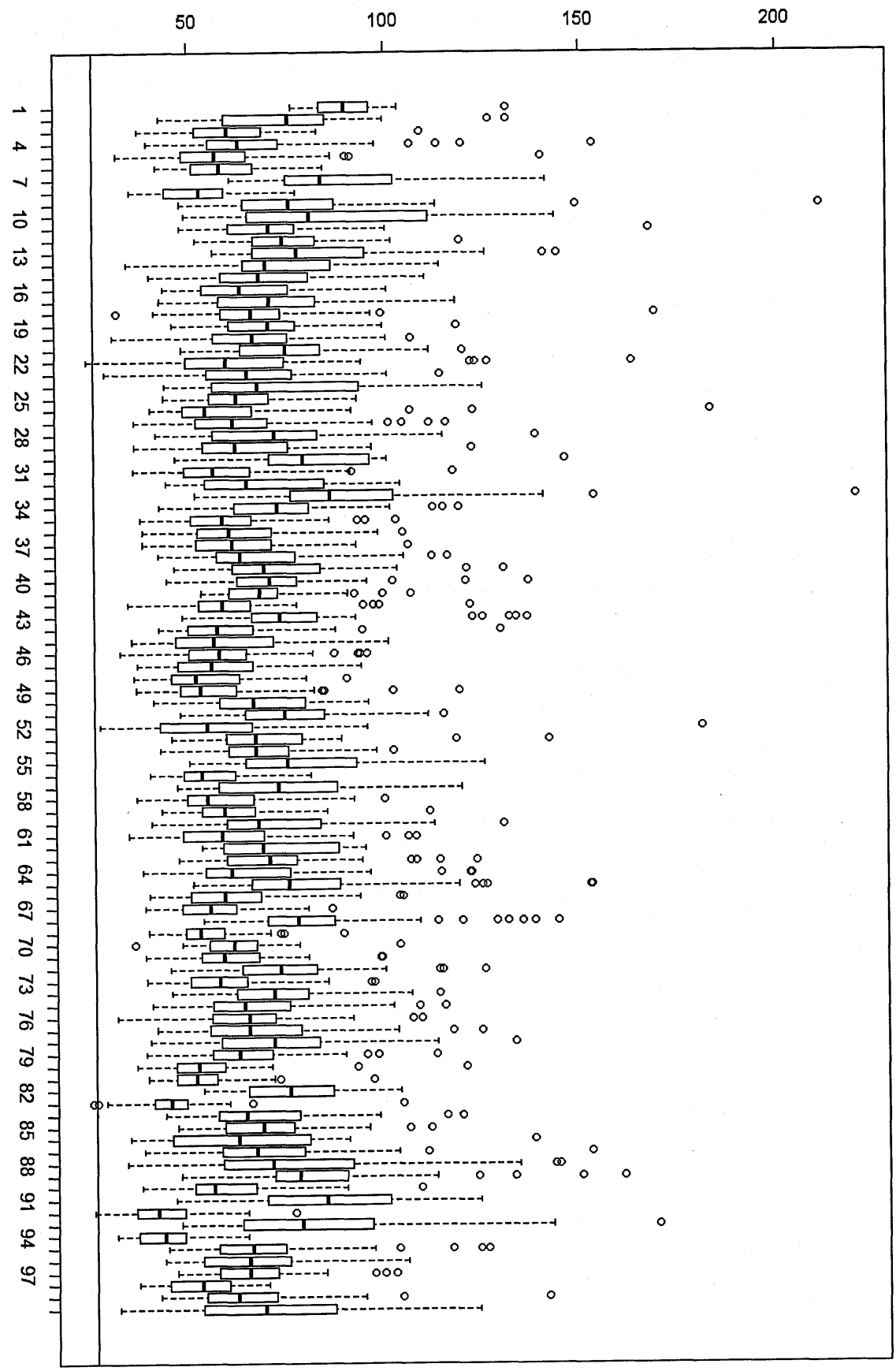


Fig. 6.1 Boxplot for the original times of the 100 mean courses

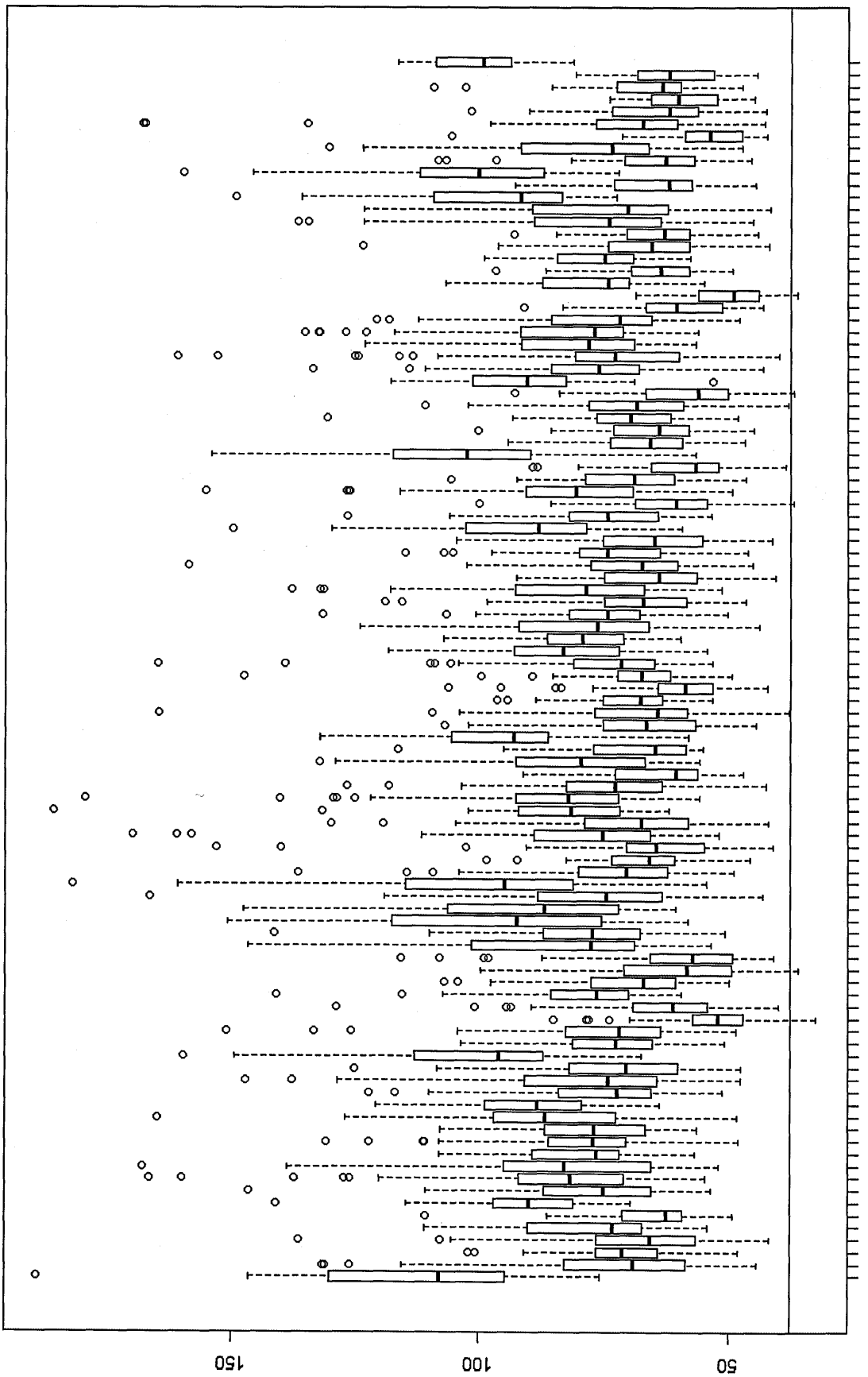
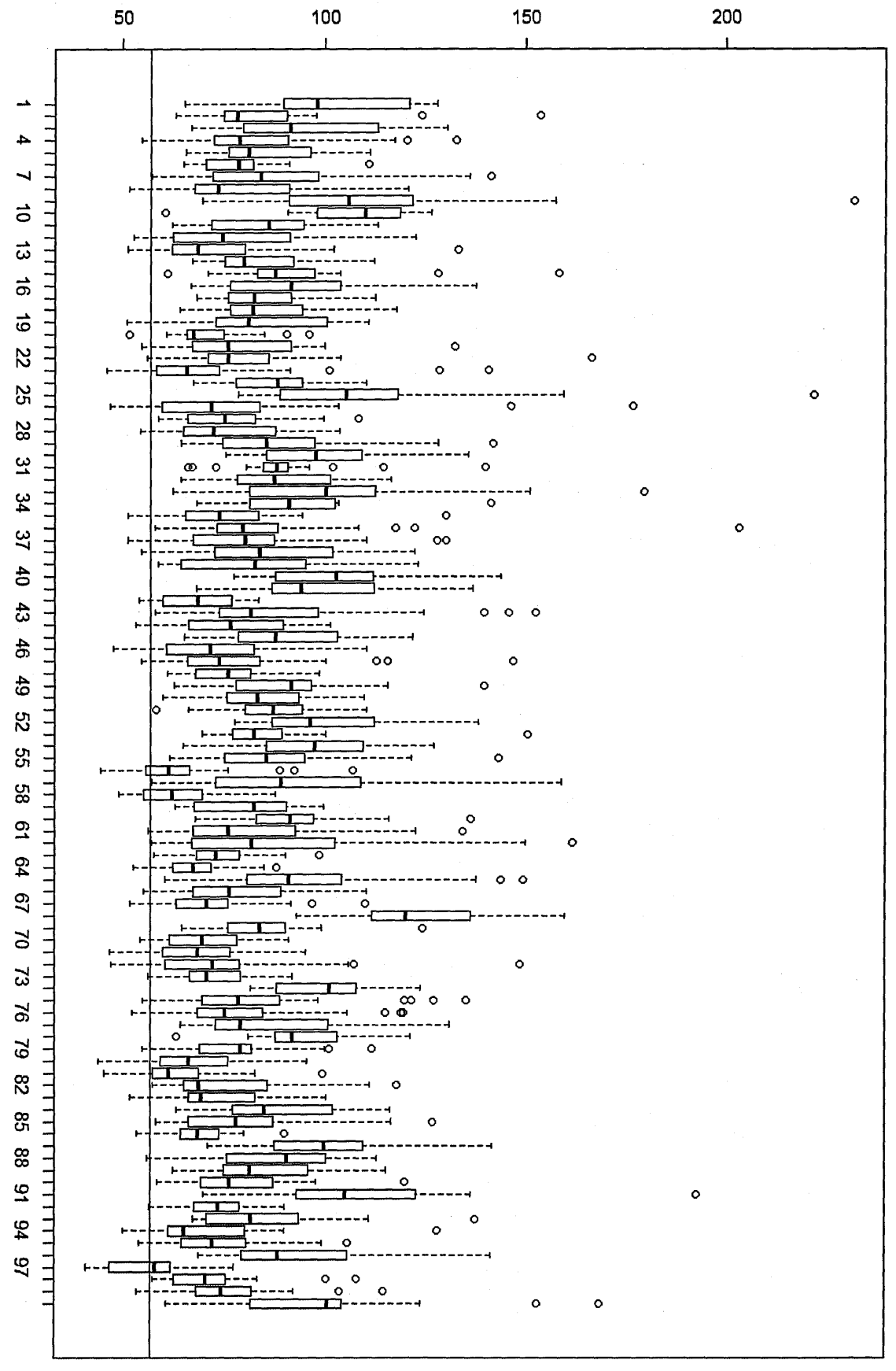


Figure 6.9. Boxplot for the original times of the 100 11


50
100
150
200
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97

6.2 Course measures

According to BOF the technical difficulty of a course is based on the skills needed to complete the course. A level 5 course should have significant route choices and legs of different lengths forcing regular changes in technique. As for the control points, they should be as few as possible and far from obvious relocating features. Also according to the rules, errors by orienteers in a course with level 5 of technical difficulty can result in a large time loss (Appendix B to Rules of Orienteering).

As mentioned in the previous section, the original times and additional information published on the BOF results web page are not enough to analyse a course technical difficulty level. For these reasons we propose the use of the orienteers' performance measures described in Chapter 5 to study the courses' technical difficulty.

An orienteer's performance measures are calculated for all the competitors in each course, then the information is aggregated by course obtaining the following course performance measures:

- Speed: Mean of the estimated orienteer's speeds for the course
- Coefficient of variation for the speed: Robust scale estimator Sn(orienteer's speed) divided by the mean course speed.
- Speed consistency: Mean of the orienteers' speed consistency measures for the course
- Coefficient of variation for the speed consistency: Robust scale estimator Sn(orienteer's speed consistency) divided by the mean course speed consistency.

- Time lost: Mean of the estimated time lost in mistakes of all the orienteers in the course. This considers only the legs where a mistake was detected through the weights.

- Expected time to complete the course: Estimation of the mean time needed for an average orienteer to finish the course considering zero time lost.

- Time lost vs time to complete the course: Proportion of the estimated time lost in mistake against the expected time to complete the course. This is:

$$\frac{\text{mean time lost in the course}}{\text{mean time to complete the course}} \tag{6.1}$$

Figures 6.4, 6.5 and 6.6 present the histograms of the course performance measures (course speed, course speed consistency, course time lost and expected time to complete the course) for the three different analysed colours. Some of the histograms show presence of atypical observations, so in the analysis we will refer back to the data and identify those atypical courses.

Figure 6.4 presents the results for the green courses. These plots show that some courses have values on the performance measures that differ from the rest. For example for the course speed there are two courses on the lower end of the distribution with a speed lower than 55 metres per minute. Those courses are the $1st$ and $68th$ courses. On the other end of the distribution two courses ($94th$ and $83rd$) have speeds higher than 110 metres per minute. Then under the assumption that all the courses were run by groups of orienteers with similar abilities, the results suggest that the $94th$ and $83rd$ courses were more runnable than the rest. The speed consistency measure has an overall mean value of 0.21. The histogram of this measure suggests one outlier at the upper end of the distribution, this is the $39th$ course

6.2 Course measures

According to BOF the technical difficulty of a course is based on the skills needed to complete the course. A level 5 course should have significant route choices and legs of different lengths forcing regular changes in technique. As for the control points, they should be as few as possible and far from obvious relocating features. Also according to the rules, errors by orienteers in a course with level 5 of technical difficulty can result in a large time loss (Appendix B to Rules of Orienteering).

As mentioned in the previous section, the original times and additional information published on the BOF results web page are not enough to analyse a course technical difficulty level. For these reasons we propose the use of the orienteers' performance measures described in Chapter 5 to study the courses' technical difficulty.

An orienteer's performance measures are calculated for all the competitors in each course, then the information is aggregated by course obtaining the following course performance measures:

- Speed: Mean of the estimated orienteer's speeds for the course
- Coefficient of variation for the speed: Robust scale estimator Sn(orienteer's speed) divided by the mean course speed.
- Speed consistency: Mean of the orienteers' speed consistency measures for the course
- Coefficient of variation for the speed consistency: Robust scale estimator Sn(orienteer's speed consistency) divided by the mean course speed consistency.

- Time lost: Mean of the estimated time lost in mistakes of all the orienteers in the course. This considers only the legs where a mistake was detected through the weights.

- Expected time to complete the course: Estimation of the mean time needed for an average orienteer to finish the course considering zero time lost.

- Time lost vs time to complete the course: Proportion of the estimated time lost in mistake against the expected time to complete the course. This is:

$$\frac{\text{mean time lost in the course}}{\text{mean time to complete the course}} \tag{6.1}$$

Figures 6.4, 6.5 and 6.6 present the histograms of the course performance measures (course speed, course speed consistency, course time lost and expected time to complete the course) for the three different analysed colours. Some of the histograms show presence of atypical observations, so in the analysis we will refer back to the data and identify those atypical courses.

Figure 6.4 presents the results for the green courses. These plots show that some courses have values on the performance measures that differ from the rest. For example for the course speed there are two courses on the lower end of the distribution with a speed lower than 55 metres per minute. Those courses are the $1st$ and $68th$ courses. On the other end of the distribution two courses ($94th$ and $83rd$) have speeds higher than 110 metres per minute. Then under the assumption that all the courses were run by groups of orienteers with similar abilities, the results suggest that the $94th$ and $83rd$ courses were more runnable than the rest. The speed consistency measure has an overall mean value of 0.21. The histogram of this measure suggests one outlier at the upper end of the distribution, this is the $39th$ course

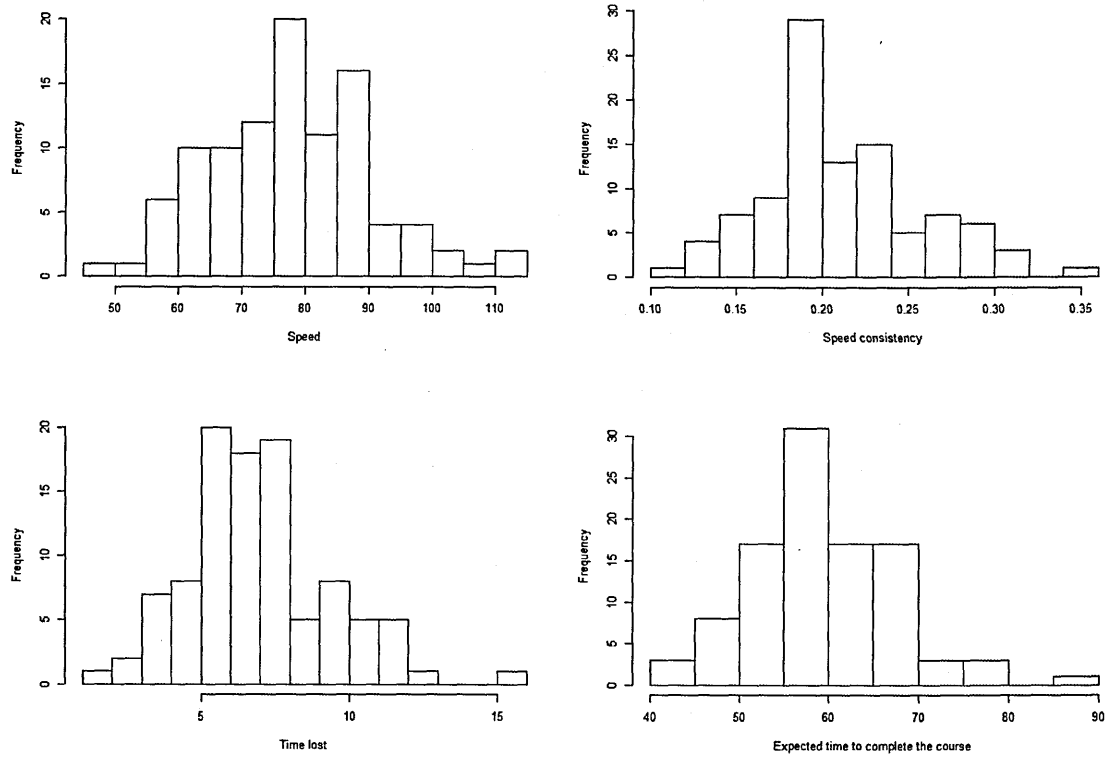


Figure 6.4: Histograms of the course performance measure of the 100 green courses.

which was the hardest to navigate with a score of 0.35. The estimated time lost for a green course is 7 minutes, however the histogram shows that the most difficult had a time lost of 15.32 minutes, more than twice the colour mean time lost, this corresponds to the $33rd$ course. The histogram of the completing time measure suggest the presence of one outlier (the $1st$ course) with larger than usual times.

Figure 6.5 presents the results for the blue courses. The course speed presents two courses with the lowest speeds considerably below the overall mean of 97 metres per minute. These are the first and $68th$ courses, which were also the two courses with the lowest speeds for the green courses. These two courses are also the ones with the largest times to complete the course with 94 and 93 minutes respectively, considerably above the overall mean time to complete a blue course of 68 minutes. For the completing time measure the atypical value seems to be on the lower end,

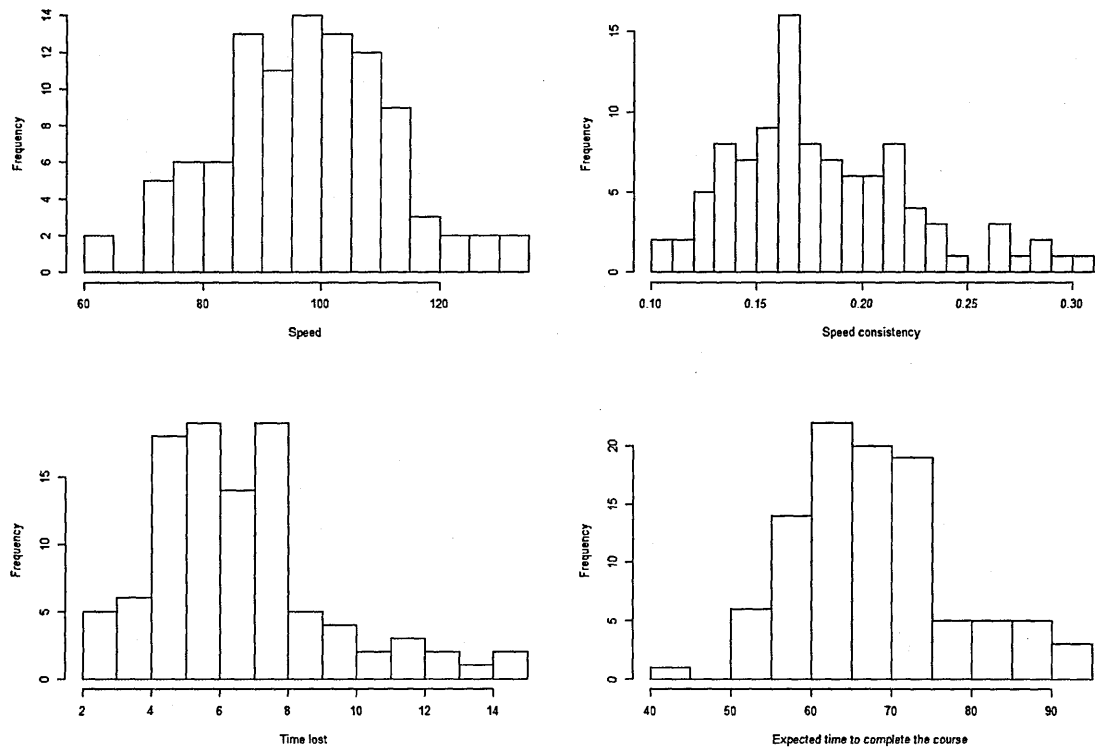


Figure 6.5: Histograms of the course performance measure of the 100 blue courses.

a course ($81st$) with a completing time of 45 minutes, suggesting this was an easy blue course. The histogram of the speed consistency and time lost do not show any course that could be markedly different from the rest of the courses. But both distributions show a heavy right hand tail.

Figure 6.6 presents the results for the brown courses. The overall speed for a brown course is 115 metres per minute, the mean time lost is 6 minutes and the mean time to complete the course is 77 minutes. As for the speed consistency measure, orienteers running a brown course have a mean score of 0.16. The histograms show that some courses have values in their performance measures that could be consider different from the rest as they lie far from the others. For example for the course speed there is one observation that lies on the lower end of the distribution, this speed of 70 metres per minute corresponds to the $68th$ course and it suggests that

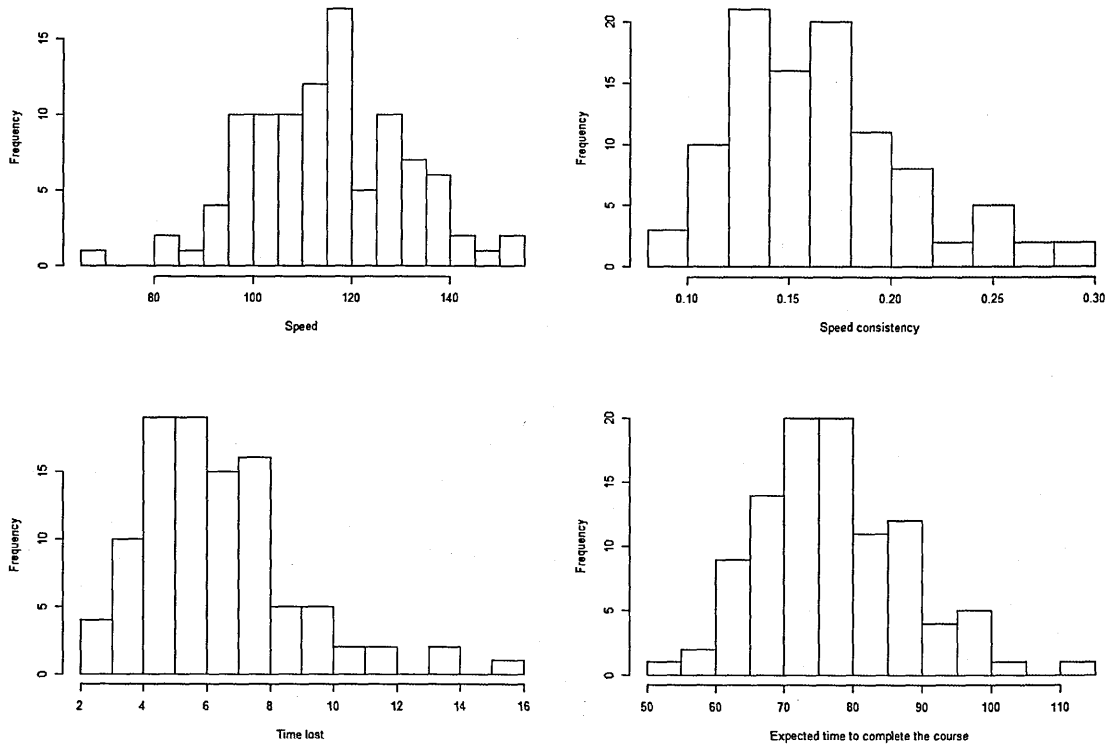


Figure 6.6: Histograms of the course performance measure of the 100 brown courses.

this course might have been harder than the average brown course. The $68th$ course was highlighted also as with the lowest speeds for the green and blue courses, this might suggests that the area where this event took place was not very runnable compared with the other events. On the top end of the time lost distribution we observed three courses (62, 91 and 33) with time lost over 13 minutes. For the completing time distribution two possibly atypical values on the top are observed, they correspond to the courses numbers 25 and 68 with 103 and 112 minutes respectively.

Comparing the results of the speed consistency scores between the three colours shown in Table 6.3 we see that the green courses have the greatest score, following by the blue colour and the brown with the smallest score. The mean time lost is very similar for the three colours. The mean time to complete the courses increases as the length increases, however the speeds also increase as the course becomes longer.

	speed	consis	t_ lost	t_ comp
GREEN	77.68	0.21	6.92	68
BLUE	97.27	0.18	6.53	76
BROWN	115.15	0.16	6.23	84

speed: Speed
consis: Speed consistency
t_ lost: Time lost
t_ comp: Expected time to complete the course

Table 6.3: Mean of the 100 courses performance measures by colour.

This analysis shows that the course speed, time lost and speed consistency differ from course to course. Even though the variation on the courses performance measures cover a wide range there is enough evidence to suggest that it would be possible to identify courses that differ from the rest (that are atypical).

6.2.1 Correlation coefficients

The results presented in the previous section show that some courses' performance measures are noticeably different from the mean values for their corresponding colours, so the following step is to study the interaction between these course performance measures. The objective is to reduce the information from the seven performance measure variables into one variable that represents the technical difficulty of the course.

The first analysis is to calculate the correlation coefficients between these variables. The correlations for the green courses presented on Table 6.4 shows that the time lost and time to complete the course are highly correlated with the speed, the speed consistency and between each other.

Similar high correlations were observed for the blue and brown courses. From these results it could be said that the time lost, the time to complete the course and the

proportion between this two (time lost ratio) are highly correlated with the speed and speed consistency variables. In the following subsection we analyse how this variables can be grouped based on the correlation observed.

	speed	consis	spd_cv	con_cv	t_ lost	t_ comp	lost/comp
speed	1						
consis	-0.47	1					
spd_cv	-0.18	0.04	1.00				
con_cv	-0.21	0.21	0.03	1.00			
t_ lost	-0.53	**0.85**	0.07	0.28	1.00		
t_ comp	**-0.77**	0.49	0.09	0.21	**0.65**	1.00	
lost/comp	-0.31	**0.82**	0.08	0.27	**0.93**	0.35	1.00

consis: Speed consistency
spd_cv: Speed coefficient of variation
con_cv: Speed consistency coefficient of variation
t_ lost: Time lost
t_ comp: Expected time to complete the course
lost/comp:Time lost over expected time to complete the course

Table 6.4: Correlation coefficients for green course variables.

6.2.2 Principal components analysis

Continuing the idea of reducing the course performance measure variable we assess a dimension reduction through a principal component analysis. Based on the correlation matrix in Subsection 6.2.1 we performed with the course performance measures (speed, speed coefficient of variation, speed consistency, speed consistency coefficient of variation, time lost, expected time to complete the course and time lost ratio). The results for each colour are presented in the following tables.

Table 6.5 presents the loadings of each course variable in the seven different estimated components, if the magnitude of the loading is smaller than 0.1 it does not appear in the table. All the components are a combination of at least three of the variables. However, the loadings suggest that components 3 and 4 have a predominant variable; the speed coefficient of variation for the third component and the speed consistency coefficient of variation for the fourth component.

Course variables	**Components**						
	1	2	3	4	5	6	7
Speed	0.37	0.50	-0.31	-0.10	0.61	0.37	
Speed consistency	-0.46	0.25	-0.13	0.14	-0.41	0.72	
Speed cv		-0.58	-0.79	-0.14			
Speed consistency cv	-0.20		0.17	-0.96			
Time lost	-0.50	0.19			0.28	-0.23	0.75
Expected time to complete the course	-0.40	-0.40	0.37	0.15	0.62	0.27	-0.28
Time lost/Time to complete	-0.44	0.39	-0.30			-0.46	-0.59
Importance of components:							
Cumulative Proportion of Variance	0.51	0.68	0.81	0.94	0.97	0.99	1.00

Table 6.5: Principal components analysis for green courses.

The first 4 components cumulate 94% of the variability of the data. But the first component alone cumulates 51% of this variability. This suggest a possible reduction in the dimensionality of the data, from seven variables to one component. This first component can be seen as a weighted combinations of the following five course variables: speed, speed consistency, time lost, expected time to complete the course and time lost ratio.

Course variables	**Components**						
	1	2	3	4	5	6	7
Speed	0.41	-0.34	-0.34		0.71	0.30	
Speed consistency	-0.44	-0.31	-0.19	0.19	-0.30	0.74	
Speed cv	-0.18	-0.31		-0.93			
Speed consistency cv		0.52	-0.81	-0.25			
Time lost	-0.50			0.12	0.35	-0.24	0.74
Expected time to complete the course	-0.37	0.55	0.35		0.50	0.32	-0.30
Time lost/Time to complete	-0.46	-0.34	-0.24	0.16	0.15	-0.46	-0.60
Importance of components:							
Cumulative Proportion of Variance	0.53	0.67	0.81	0.94	0.98	0.99	1.00

Table 6.6: Principal components analysis for blue courses.

The results of the principal components analysis for the blue courses are presented in Table 6.6. The seven components are a combination of at least three of the variables, but the loading values suggest that components 3 and 4 have a predominant variable. These variables are: 1) The speed consistency coefficient of variation for

the third component. 2) The speed coefficient of variation on the fourth component.

The table also shows that the first component alone cumulates 53% of the variability of the data. Compared with the results in Table 6.5 for the green course, we see that for the first component if only loadings over 0.20 are considered, then these two colour coded courses have very similar first component.

Course variables	**Components**						
	1	2	3	4	5	6	7
Speed	0.31	-0.58		-0.11	0.73	-0.12	
Speed consistency	-0.46	-0.27	0.17	-0.35		0.75	
Speed cv	-0.20	-0.17	0.58	0.76	0.10		
Speed consistency cv	-0.12	-0.17	-0.78	0.52		0.28	
Time lost	-0.56		-0.11		0.14	-0.36	-0.72
Expected time to complete the course	-0.25	0.67			0.65		0.26
Time lost/Time to complete	-0.51	-0.30			-0.12	-0.45	0.64
Importance of components:							
Cumulative Proportion of Variance	0.43	0.64	0.80	0.91	0.96	0.99	1.00

Table 6.7: Principal components analysis for brown courses.

Table 6.7 presents the loadings of the course variables for the estimated components on the brown courses analysis. The results show that except for the last component, all the components are a combination of at least four of the variables, and the loading values do not suggest that any of the components has a predominant variable. The third and fourth component are mainly based on the speed and speed consistency coefficients of variation. Component 5 is speed, expected time to complete the course and the proportion between this last one and the time lost. Components 6 and 7 are mainly speed consistency and time lost respectively.

The table also shows that the first component alone cumulates 43% of the variability of the data. Compared with the results for the green and blue course, we see that this first component is formed by the same variables (speed, speed consistency, time

lost, expected time to complete the course and time lost ratio) if only loadings over 0.20 are considered.

The principal components analysis approach to reduce the colour coded courses data suggest a possible dimensional reduction from 7 variables to 4 components, and it covers between 91% or 94% of the variance depending on the course colour. The results also shown that the first component is similar for the three colours and it explains around half of the variance in the data. Our interest is to find a technical difficulty calculation that works for all the colours. For those reasons we explore the use of a simple principal component with loadings values -1, 0 and 1. This is based in the simplification suggested by Jackson (1991) where drastic rounding of the loading values of the principal components analysis is used to obtain simple approximations of the components.

Using the loadings for the first component obtained from the PCA we define the loadings for the simple principal component. Assign value of zero to loadings between -0.20 and 0.20, if the loading is greater than 0.20 it will be rounded to 1 and if it is smaller than -0.20 it will be rounded to -1.

Then we have that for the three colour courses the simple principal component will be of the form:

Course variables	**Component**
Speed	1
Speed consistency	-1
Speed cv	0
Speed consistency cv	0
Time lost	-1
Expected time to complete the course	-1
Time lost/Time to complete	-1

Table 6.8: Simple principal components analysis

This simple principal component explains for the green course 50% of the variance, which is a good approximation of the first principal component. For the blue and brown courses the variance explained by the simple principal component is 51% and 40% respectively. These are also good approximation to the results for the first component in the PCA. This analysis suggests that the simple principal component presented in Table 6.8 can be used as a measure of the course technical difficulty.

6.3 Technical difficulty measure

As mentioned at the beginning of this chapter the course technical difficulty is in terms of both the physical and navigational skills the orienteers should have to complete the course. This section presents two approaches to measure the course technical difficulty based on the results of the simple principal component analysis in Subsection 6.2.2. One method is just to define the difficulty measure as the simple principal component found. This component is based on the speed, speed consistency, time lost, expected time to complete the course and time lost ratio. A different approach will be to construct the course technical difficulty measure using ranks on the variables of the simple principal component.

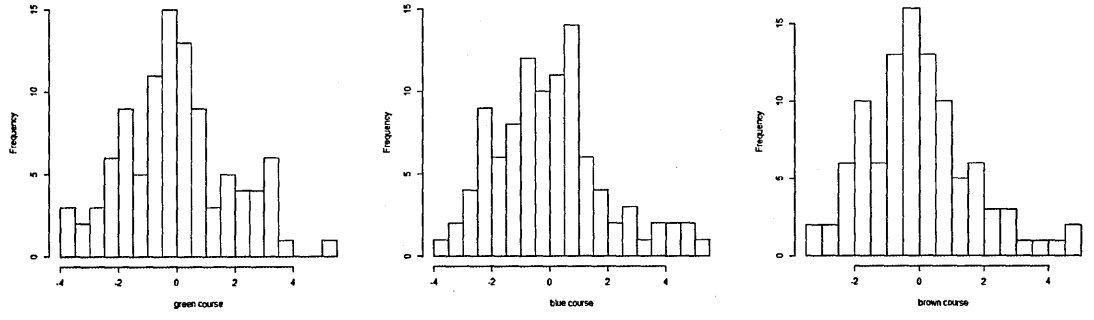
6.3.1 Simple principal component

This course technical difficulty measure is defined with the simple principal component in Table 6.8. Then a course difficulty can be calculated with the following equation:

$$\begin{aligned}\text{course technical difficulty} = \frac{-1}{\sqrt{5}}(&\text{speed} - \text{speed consistency} - \text{time lost} \\ &-\text{expected time to complete the course} - \frac{\text{time lost}}{\text{time to complete the course}})\end{aligned} \tag{6.2}$$

with all the variables standardized.

The histograms in Figure 6.7 present the distribution of this difficulty measure for each colour. The plot for the green courses suggest that one of the courses has an atypical difficulty. And for the blue and brown courses the histograms show that the distributions are heavy on the right hand side tail.



(a) (b) (c)

Figure 6.7: Density technical difficulty based on simple principal component for each colour.

The high outlier on Figure 6.7a corresponds to the $33rd$ course, which in the analysis done in Section 6.2 was detected as a green course with a higher than usual time lost. We also detected the first and $68th$ courses with very low speed, this courses obtained a difficulty measure of 3.38 and 2.38 respectively. The $39th$ course, had a high value on the speed consistency measure, meaning that the speeds on that course were not very consistent, and the difficulty value is detecting that as well giving this course a 2.98 difficulty level. On the other end of the difficulty scale we have courses $83rd$ and $94th$ with -3.10 and -3.89 of difficulty, this courses show high values of speed. This suggest that this course technical difficulty is detecting the atypical courses we observed in Section 6.2.

For the blue courses we have that the difficulty of the first and $68th$ courses are 5.42 and 4.02 respectively. In Section 6.2 these two courses presented atypical speeds and times to complete the course. However the $81st$ course has a difficulty of only

-1.82, and we observed this course had a significantly shorter time to complete the course in comparison with the other blue courses. On the other end of the difficulty scale, the $49th$, $94th$ and $22nd$ courses have a difficulty value under -3. We observed that course number 94 was also detected on the lower end of the difficulty scale for the green course.

On the brown courses, the three course with difficulty greater than 4 observed in Figure 6.7c correspond to course $33rd$, $91st$ and $68th$. The first two had the highest time lost for brown courses, and the $68th$ course had atypical low speed and the longest time to complete the course. In Section 6.2 we observed that the $62nd$ course had a time lost very similar to the $91st$ course, but the difficulty measure is giving a value of 3.52 which is large and positive but does not seem atypical in the distribution. Also the $25th$ course had a long time to complete the course and its correspondent difficulty was only 0.87. Similar to the green and blue course, the $94th$ course is in the lower tail of the difficulty distribution with a value of -3.09.

This course technical difficulty seems to provide a sensible measure for the course difficulty based on the speed, speed consistency, time lost, expected time to complete the course and time lost ratio. However if the combination between this 5 variables is such that the course technical difficulty balances the high values, then the measure will not be detecting courses that are different from the rest. That was the case for the $81st$ blue course and $25th$ brown course.

6.3.2 Technical difficulty ranking

Following the results in the previous subsection we now consider the construction of a technical difficulty ranking based on assigning ranks to the variables of speed,

speed consistency, time lost and expected time to complete the course. Table 6.4 shows that the time lost ratio is highly correlated with the time lost variable, so the information that can provide on the ranks is similar and for that reason the ratio variable was dropped out for this ranking. These four variables are part of the simple principal component analysis, and the variance associated with a simple component with only these 4 variables is 41%, 41% and 32% for the green, blue and brown courses respectively. It is considered that the course speed and time to complete the course can be used to estimate the physical difficulty of the course, where higher speed and low times are seen as low physical difficulty. And the speed consistency score and time lost are used to estimate the navigational difficulty of the course.

The difficulty ranking proposed is based on a categorisation of the four selected course variables. Each variable will have three categories low, medium and high and the levels were defined using the quantiles (33% and 66%) of the variables empirical distributions. We define the technical difficulty as the addition of both the physical and navigational difficulties.

A course with small speed will be in the high category of the physical difficulty. In contrast the expected time to complete the course has direct relationship with the physical difficulty, so courses with large values of expected time to complete will be ranked in the highest category. Also the speed consistency scores and time lost have a direct relationship with the navigational difficulty, so courses with high speed consistency scores or large time lost will tend to be ranked in the highest category.

The empirical distributions of course variables are based on the results for the same events analysed in Sections 6.1 and 6.2. The three categories for each variable are presented in Tables 6.9 and 6.10.

		Physical difficulty level		
		high (3)	medium (2)	low (1)
speed	**green**	≤71.87	>71.87 & ≤ 83.34	> 83.34
	blue	≤90.66	>90.66 & ≤102.70	>102.70
	brown	≤108.48	>108.48 & ≤119.86	>119.86
		low (1)	medium (2)	high (3)
time to complete the course	**green**	≤55.86	>55.86 & ≤ 62.30	> 62.30
	blue	≤62.86	>62.86 & ≤70.82	>70.82
	brown	≤71.03	>71.03 & ≤79.76	>79.76

Table 6.9: Physical difficulty levels for the colour courses

Table 6.9 presents the cut points for the course variables that measures the speed of an average orienteer running the course and the expected time to complete the course. It is important to remember that both variables are calculated without the time lost on navigational mistakes, so the speed estimates how runnable the course was.

The speed values for each colour shown in Table 6.9 and histograms in Figures 6.4, 6.5 and 6.6 suggest that this speed measure is influenced by the fitness of the orienteers as the speeds for the brown courses are faster in relation to the green and blue courses. Similarly the orienteers running blue courses tend to be faster than the orienteeers on green courses. This suggest that the difference between these three colour courses might not be only their lengths. It is also possible that if the physical difficulty of a course is one of the factors that influences orienteers' course selection, faster and so fitter runners will tend to choose a brown course.

As mentioned at the beginning of this chapter, according to the BOF guidelines the green, blue and brown courses should have a level 5 of technical difficulty. For this reason it is expected that the values of the variable measuring the navigational difficulty do not differ much between colours. However because these variables are constructed from an orienteer's performance measures the course technical difficulty will depend on the competitors' perceptions of the course, so the navigational difficulty as we propose depends on the skills of the orienteers running each colour. The ranks for the navigational factors are presented in Table 6.10.

		Navigational difficulty level		
		low (1)	medium (2)	high (3)
speed consistency	**green**	$\leq$0.19	>0.19 & $\leq$0.22	>0.22
	blue	$\leq$0.16	>0.16 & $\leq$0.19	>0.19
	brown	$\leq$0.14	>0.14 & $\leq$0.17	>0.17
time lost	**green**	$\leq$5.63	>5.63 & $\leq$7.39	>7.39
	blue	$\leq$5.25	>5.25 & $\leq$7.09	>7.09
	brown	$\leq$5.00	>5.00 & $\leq$6.93	>6.93

Table 6.10: Navigational difficulty levels for the colour courses

An explanation for the decreasing effect observed across the colours is again the orienteers' course self selection. In term of the navigational skills most of the competitors on a brown course are experienced and confident orienteers, in contrast with the fact that non experienced orienteers will be encouraged to run a green course to gain experience and then move up to blue course. Comparing the three colours, the navigational difficulty looks more consistent than the physical difficulty, which was suspected as the three courses have the same technical difficulty level according to the BOF guidelines. This suggest that these navigational difficulty measures are reflecting the designed course technical difficulty.

After calculating the categories each course is assigned its corresponding ranking on the four course variables. Rank equal to 1 for a low category, rank equal to 2 for a

medium category and rank equal to 3 for a high category. Finally the sum of the four ranks will be the final course technical difficulty. This difficulty scale goes from a value of 4 for very easy courses to 12 for very hard courses.

The example on Table 6.11 shows how the course technical difficulty is estimated. The example consists of 3 courses chosen from the 100 green courses to demonstrate the technique.

	Course variables				Phy		Nav		Technical
Course	speed	t_ comp	consis	t_ lost	r1	r2	r3	r4	**Difficulty**
1	45	85.30	0.28	8.60	3	3	3	3	12
2	62	68.30	0.18	5.03	3	3	1	1	8
3	92	54.97	0.16	4.70	1	1	1	1	4

consis: Speed consistency
t_ lost: Time lost
t_ comp: Expected time to complete the course
Phy: Physical difficulty
Nav: Navigational difficulty

Table 6.11: Example of technical difficulty with three green courses

According to the difficulty scale the first course was a very difficult course, the second an average green course and the third a very easy course. This example also illustrate that courses 1 and 2 have the same high level on the physical difficulty and in this case is the difference in their navigation difficulty the cause of the second course being an average course and the first a hard course. The ranks on Table 6.11 suggest that course number two has a navigational difficulty equals to 2 (r3+r4), caused by both a low speed consistency score and small time lost.

Figure 6.8 presents the technical difficulty for all the courses in the analysis. The scatter plot shows how consistent is the technical difficulty in an event.

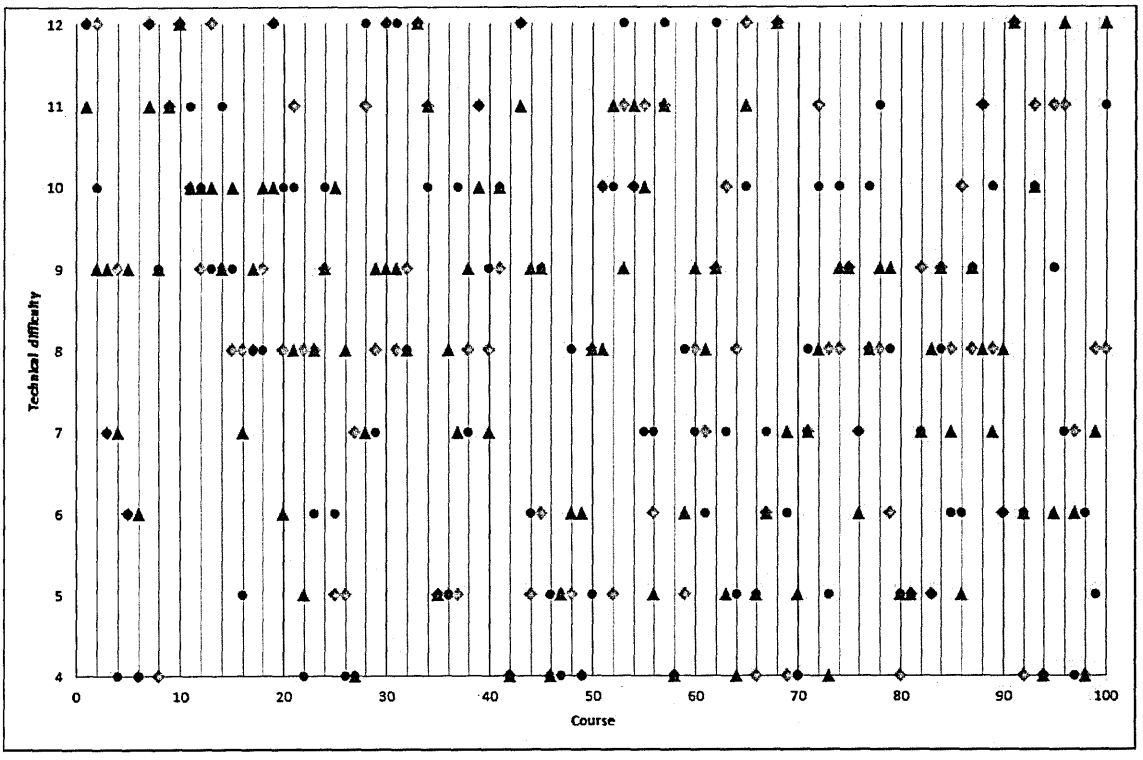


Figure 6.8: Technical difficulty of the 100 courses in the three colours.

Figure 6.8 shows that the technical difficulty across courses in an event can be very variable. There are events for which the course technical difficulties are identical for the three colours, in this analysis there are 11 courses with this characteristic and two examples are the $10th$ and $42nd$ event. But there are others like the $28th$ event which seems to have a difficult green and blue courses and an average brown course, or the $64th$ event with average green course an easy blue and brown courses. From this figure we also observed that the courses number 10, 33, 68 and 91 have consistently high difficulty level in all the colours. On the other end the courses number 42, 58 and 94 have very low technical difficulty for the three colours.

The following figures present the results of the course technical difficulty for each colour. To help the visualization of the results the plots use the physical and navigational difficulties as the axes. The navigational difficulty value is the sum of the ranks assigned to the two course variables, as shown in the previous example. Then

the circles plotted correspond to the number of courses that have an x value of physical difficulty and a y value of navigational difficulty. Finally the diagonal lines represent the technical difficulty, and this goes from 4 on the down left hand side of the graph to 12 on the upper right hand side.

To help the interpretation of the technical difficulty ranking, values of 11 and 12 are considered as highly technically difficult (or hard) courses, course with values of 4 and 5 are consider as with low technical difficulty (or easy) courses and values between 6 and 10 correspond to courses with average technical difficulty.

Figure 6.9 shows the distribution of the green courses by their technical difficulty measures. This results suggest that 25% of the courses were classified with high technical difficulty, this means that the courses have values of 11 or 12 on the difficulty scale. The figure also shows that 25% of the courses have values of 4 or 5 on the technical difficulty. Then the rest of the courses (more than 50%) have average technical difficulty.

Comparing with the observations made in Section 6.2 about the courses, we have that for the green courses, the $1st$, $33rd$ and $68th$ course have a technical difficulty of 12, the $39th$ that had a high speed consistency has a difficulty level of 11 and the $83th$ and $94th$ (the ones with highest speeds) have a difficulty of 5 and 4 respectively. In Subsection 6.3.1 was mentioned that the $33rd$ course was an outlier in the green courses simple principal component difficulty distribution, this suggests that besides being a difficult course, the $33rd$ course was much harder than the other courses.

Figure 6.9 also shows 2 courses whose physical difficulty was very low (2) but the navigational difficulty has the maximum value of 6 points. So even though this

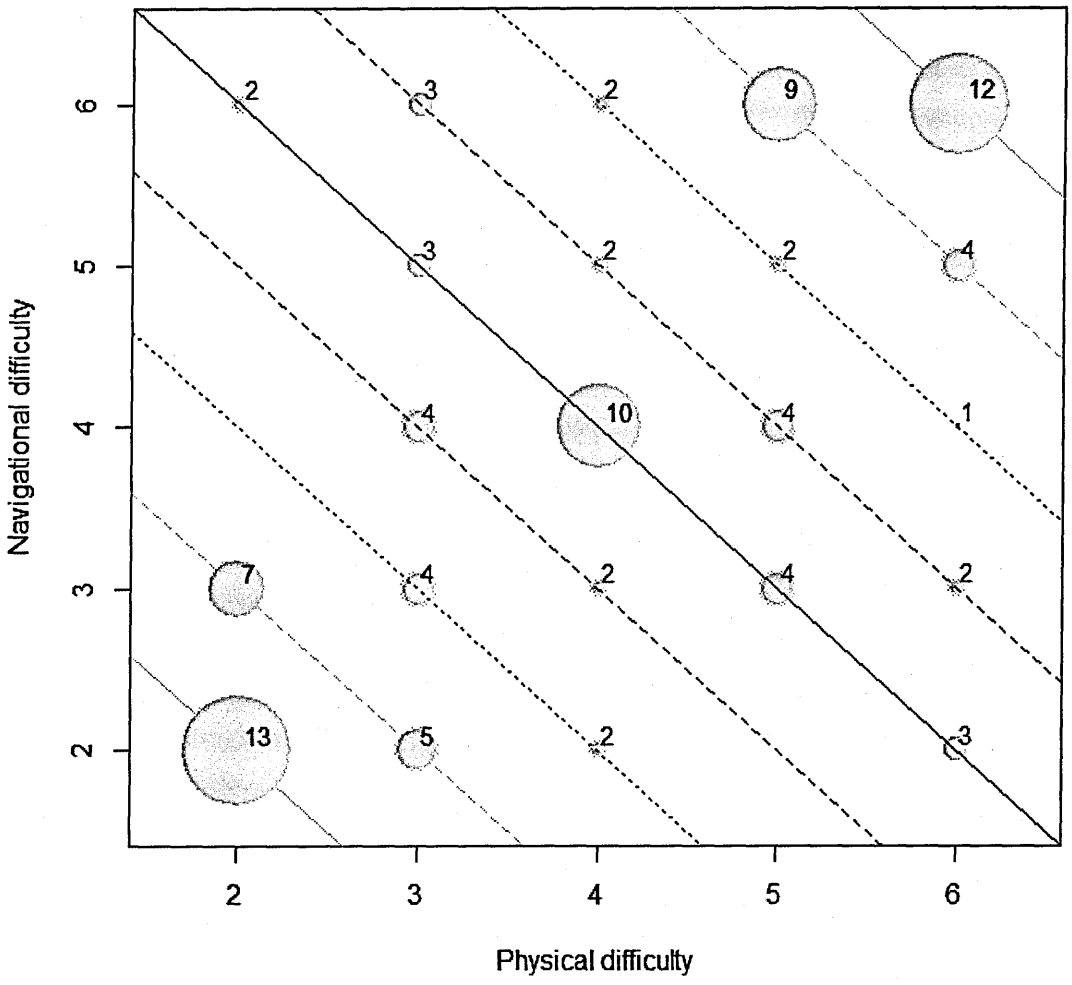


Figure 6.9: Technical difficulty of the 100 green courses.

courses have an average technical difficulty, it is possible to identify them as highly navigationally demanding. These observations correspond to the courses number 31 and 64. The same could be done to identify physically demanding courses.

The distribution of technical difficulty for blue courses is shown on Figure 6.10. The blue course have 21% of the courses on the high end of the technical difficulty and 24% on the lower values. When we compared with the results in Section 6.2 we observed that the two courses with atypical speed (events 1 and 68) have a difficulty

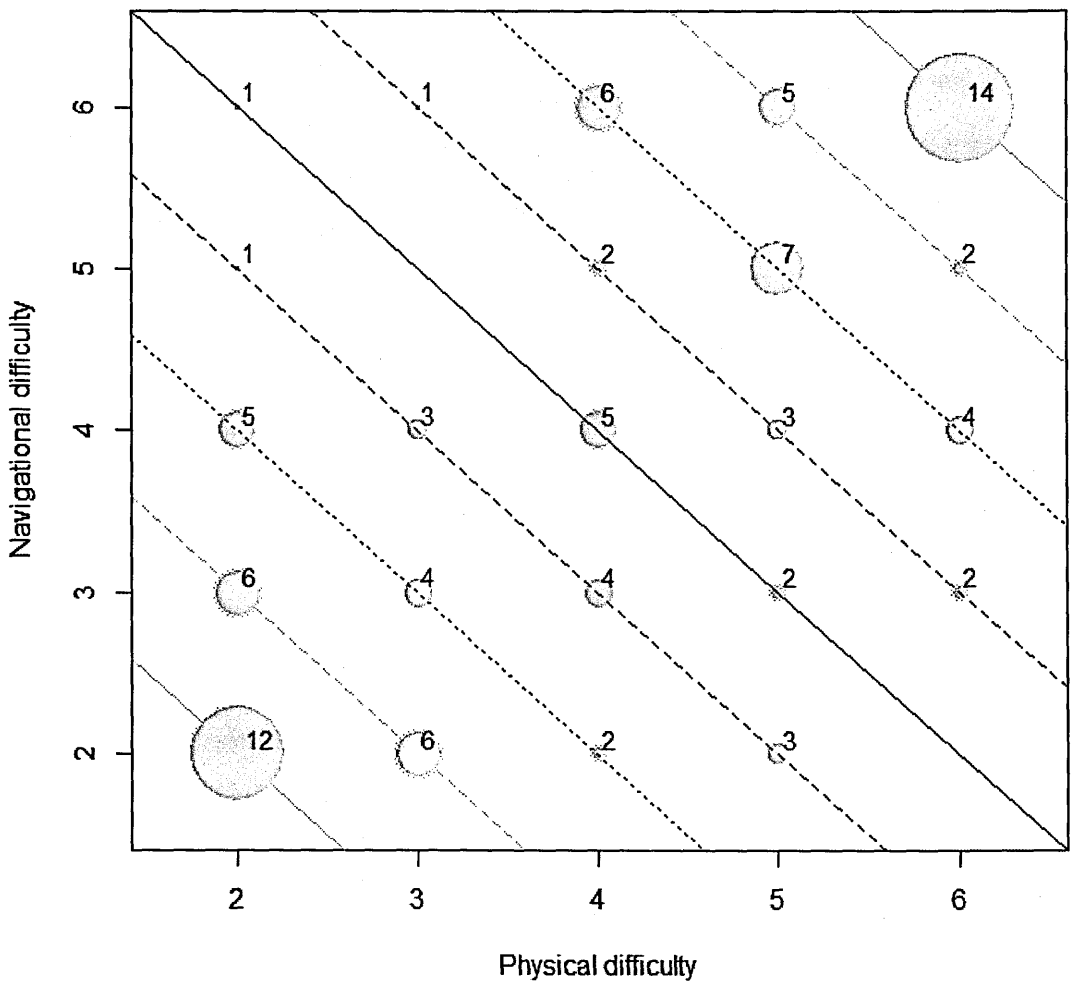


Figure 6.10: Technical difficulty of the 100 blue courses.

level of 12 and the $81st$ course, which time to complete the course was shorter than the rest has a difficulty of 5. This shows that this procedure will identify the $81st$ course as an easy course in comparison to the difficulty based on the simple principal component that did not identify this course as different.

Figure 6.10 shows that there is only one course with very low physical difficulty and high navigational difficulty. This observation corresponds to the $18th$ course, which speed is 108 metres per minute and has a time lost of 8 minutes. Based on

Figure 6.5 this course will be in the upper half of the distributions for those variables.

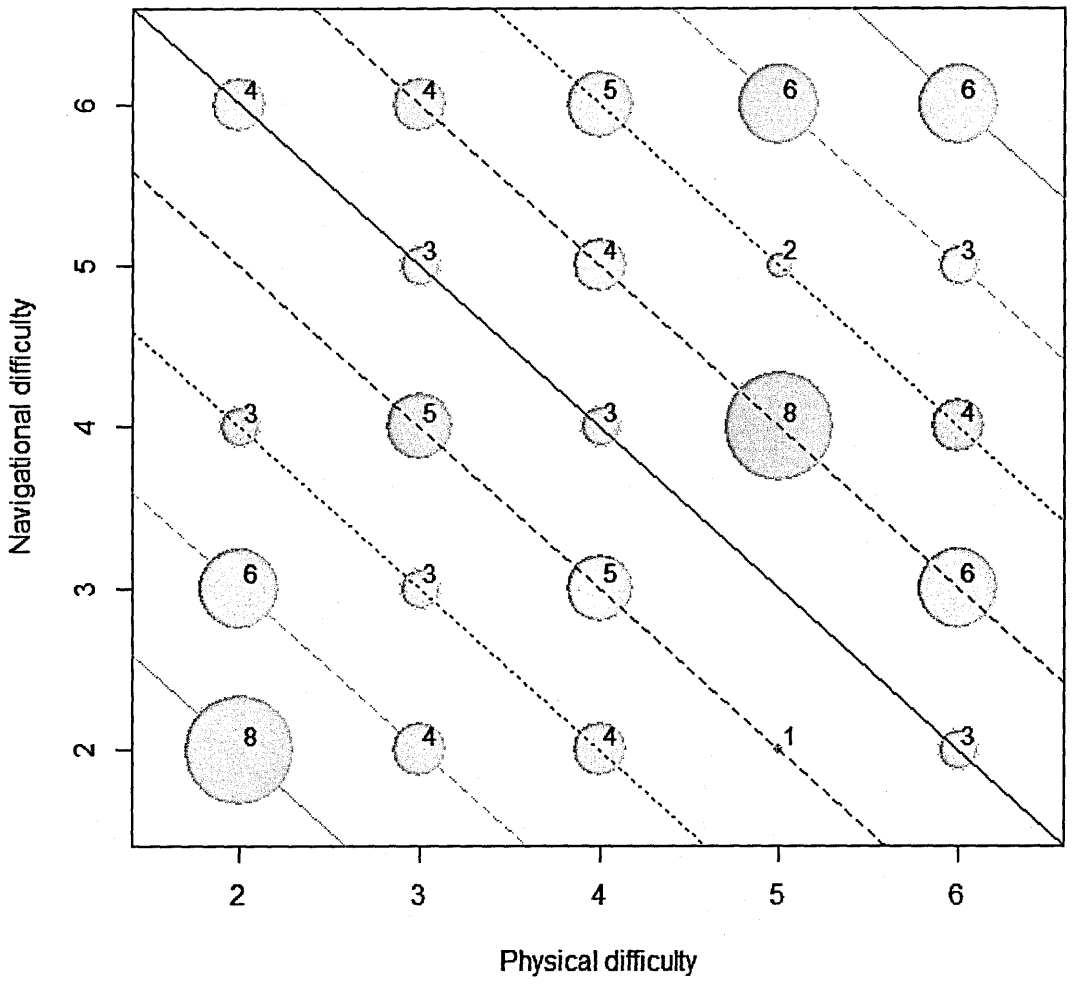


Figure 6.11: Technical difficulty of the 100 brown courses.

For the brown courses, Figure 6.11 shows that 6% have technical difficulty of 12, and 15% of the courses were classified as with high technical difficulty. On the other side of the scale, only 8% have the lowest rank of the difficulty measure, and 18% of the courses were classified as easy courses. This leaves around 67% of the brown courses with an average technical difficulty. Compared with the green and blue courses, the brown courses seem to be more spread out over the different values for both the navigational and physical difficulty.

We observed that similar to the technical difficulty measure in Subsection 6.3.1, the $33rd$, $68th$ and $91st$ courses have a high difficulty level. The $62nd$ course has an average difficulty (9 points), but the $25th$ course is now in the boundary between average and difficult with 10 points.

Figure 6.11 also shows that for brown courses there are three events with very high physical difficulty and very low navigational difficulty. And there are four events with very low physical difficulty and very high navigational difficulty. Compared to what was observed for the green and the blue courses these extreme courses seem to have a larger presence in brown courses.

The analysis presented in Sections 6.1 about the original times and in Section 6.2 about the course measures suggested that courses of the same colour might have different technical difficulties. However these differences are not in only one characteristic of the courses, such as length, number of controls, the average speed of the orienteers or the average time lost per competitor caused by navigational mistakes. The figures and the principal component analysis show that the times and performances of the orienteers have a complex correlation with the course characteristics and possibly with course factors that were not measured or even cannot be measured. This makes the construction of a technical difficulty scale challenging.

The analysis of technical difficulty presented in this chapter offers two methods to analyse a course difficulty in terms of similar courses. This comparison is based on the orienteers' performance at each course. So the following assumption is important; for each event the representation of orienteer's skills at each colour is similar.

The two difficulty measures presented in this chapter seem to detect different features of the course technical difficulty level. The simple principal component technical difficulty measures how much more easy or difficult the course was compared with the other courses (detects outliers), however it seems this method does not detect all the atypical courses. The technical difficulty ranking identifies the courses in the upper and lower end of the difficulty scale, but it does not identify outliers. This difficulty measure also allows a differentiation between a high difficulty level caused by being a less runnable area or high difficulty caused by being a course with high navigational demand. As our interest is to assign a difficulty level to the courses, we suggest the use of the technical difficulty ranking. Furthermore this difficulty measure combined with the simple principal component results could be used to detect outlier courses.

The courses used for the construction of this measures of technical difficulty considered events in different regions of the UK that took place all around the year. Because of the robustness of the sample, we are confident that these measures can be used to estimate the technical difficulty level of events outside the sample. However it is possible that certain characteristics of the courses change with time. These changes might affect the set values used in the technical difficulty ranking, so it will be desirable to continue updating the data set of events used and analyse if the values need to be modified.

Chapter 7

Another application of the model: average expenditure in Mexican households

In this chapter we discuss how the method proposed in Chapter 3 can be applied to a problem that is not related to orienteering. The algorithm developed in Chapter 3 was motivated by the specific problem of measuring the performance of orienteers in an event. In the first section of this chapter we generalize the characteristics of the orienteering problem, in particular the matrix $\mathbf{T}$. This analysis of how the data has to be in order for the method to be used, allows us to find other applications of the method. In Section 7.2 we discuss data related to expenditure in Mexican households, showing that the method can be applied in this case. Results of applying the algorithm and the comparison of those results with published results are presented in Section 7.3. The final section presents an analysis of the sensitivity of the algorithm in this application.

7.1 The method

Based on solving the problem of measuring the competitor's performance in an orienteering event, we have developed a method that approximates a matrix $\mathbf{T}$ of dimensions $n \times m$ by the multiplication of two vectors $\mathbf{a} = (a_1, a_2, ..., a_n)$ and $\mathbf{b} = (b_1, b_2, ..., b_m)$. In the case of orienteering the matrix $\mathbf{T}$ contains the times that each competitor (the rows) took to go from one check point to the next one (the columns) along the course.

The nature of the orienteering problem suggested that the matrix $\mathbf{T}$ can contain three types of atypical observations. One type are the row wise outliers, these are competitors that have larger (or smaller) times in all the columns. This kind of row represents slow (or fast) runners. The second type are column wise outliers, which correspond to particularly long (or short) legs. The third type are the element wise outliers, these outliers appear when a competitor gets lost while going from one check point to the next one, so his/her time is longer than the rest of the competitors and also relatively longer than expected based on his/her times in the other columns. In the orienteering problem we are interested in defining as outliers only the element wise outliers. Our method is specially designed to take into consideration these characteristics of the matrix. So our method is robust to atypical rows and columns, but sensitive to element wise outliers. So this suggests that the matrix $\mathbf{T}$ should have the following characteristics:

1. Each row represents an observation unit. (e.g. persons, houses, etc.)

2. Each column measures different aspects of the observation unit, but they must be measured using the same metric. (e.g. time, amount of money, metres, etc.)

3. The distribution per row might have outliers.

4. The distribution within each column might have outliers and is skewed in the same direction for all the columns. The method was developed for right-skewness, but can be easily modified to deal with left-skewness.

5. The matrix could have missing values, but the number of missing cells per row cannot be more than half of the number of columns and the number of missing cells per column cannot be more than half of the number of rows.

As a result of applying the proposed method to approximate the matrix $\mathbf{T}$ we obtain a vector $\mathbf{a}$ that summarizes all the information of the observation units in a way that each element of the vector is comparable with each other. And similarly the vector $\mathbf{b}$ will estimate a characteristic comparable between the different aspects measured. For the orienteering case we have that the vector $\mathbf{a}$ is the competitor's speed, and the vector $\mathbf{b}$ estimates the distances between check points along the course.

As we mentioned, the contribution of our method is that the estimates of the two vectors will not be influenced by the presence of missing values or outliers. Going back to the orienteering case, we have that the values in the vector $\mathbf{a}$ are the competitors' speeds regardless of getting lost or missing any check point.

7.2 Average expenditure in Mexican households

The algorithm that produces the lower rank minimization mentioned in the previous section and described in detail in Chapter 3 can be applied to household expenditure information, because this type of data is similar to the orienteering data. In this section we will analyse how the household expenditure complies with the characteristics described in Section 7.1. We use data from the National Household Income and Expenditure Survey (ENIGH) 2010, gathered by The National Institute of Statistics

and Geography of Mexico (INEGI) (INEGI, 2010).

This survey contains the information about the level and structure of the income and expenses of the Mexican households selected in the sample. In particular the survey collects expenditure information about 8 general concepts: food (FOOD), dress & shoes (DRESS), house (HOUSE), house maintenance (MAINT), health (HEAL), transport & communication (TRANS), education & hobbies (EDUC) and personal (PERS). The data represents the expenditure per household during a three month period on each of the eight general concepts. We need that the amounts between households are comparable so that we can say that each row is measuring the same thing. For this reason we divided the amounts by the number of individuals in each household this way the data can be arranged as a matrix $\mathbf{T}$ with the characteristics described in the previous section.

First we will do a basic analysis with the boxplots of the information in each of the eight concepts. As shown in Figure 7.1 all the dots outside the boxes suggest extreme skewness in the distributions and presence of outliers in the data. In Table 7.1 the number of outliers identified by the boxplot for each concept from the total of 27,655 cases in the survey is given. Health is the expenditure concept with the largest proportion of outliers, 15% of the observations.

We also observe from the boxplots in Figure 7.1 and from Table 7.1 the presence of a large number of zero values in the following concepts: dress & shoes, house maintenance, health and education & hobbies. The methodological description of the survey does not specify how the missing values for these variables were treated. So these zero values represent either zero amount expended by households or non-

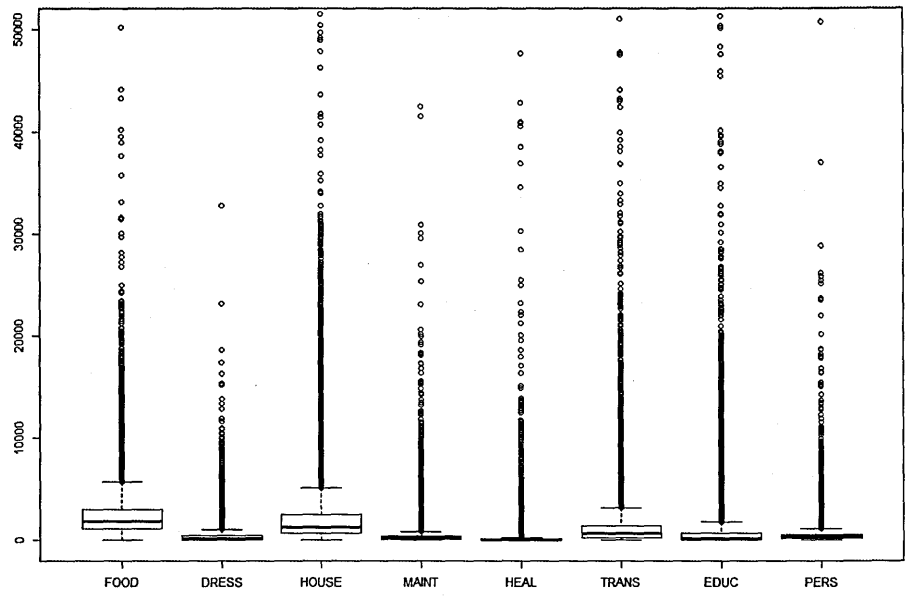


Figure 7.1: Boxplot for the eight general expenditure concepts for data form ENIGH 2010.

information provided, because there are not blanks in the original data set. Similar to the outliers case, the health concept is the one with the largest amount of zeros, with 51% of the households assigned a zero value for this concept, it is possible that some of these are actually zero amount expended in the health concept.

After identifying the presence of outliers and missing values in the data we analysed the effect of these two in the estimation of average expenditure. To measure the effect of large values, we use the outlier definition of a boxplot. That will be any observation that is higher than the third quartile by more than 1.5 times the interquartile range. We then compared the estimated mean expenditure with and without outliers for each concept.

	FOOD	DRESS	HOUSE	MAINT	HEAL	TRANS	EDUC	PERS
No. of outliers	1,791	2,327	2,433	3,284	4,151	2,089	3,268	2,535
%	6%	8%	9%	12%	15%	8%	12%	9%
No. of zeros	230	5,667	81	494	14,087	3,200	10,036	390
%	1%	20%	0%	2%	51%	12%	36%	1%

Table 7.1: Number of zeros and outliers in each concept and their corresponding percentage from the total of 27,655 cases.

	FOOD	DRESS	HOUSE	MAINT	HEAL	TRANS	EDUC	PERS
All data	$2,420	$370	$2,254	$483	$206	$1,218	$787	$514
Without outliers	$1,967	$215	$1,470	$220	$26	$751	$258	$312
Reduction	19%	42%	35%	54%	87%	38%	67%	39%

Table 7.2: Estimates of mean expenditure per household for each concept with and without outliers detected with the boxplot. The amounts are in mexican pesos.

Table 7.2 presents the effect that outliers have on the estimation of mean expenditure. Health is the concept for which estimation changes most. The estimation using the complete data suggest that households spend 206 mexican pesos per person on the health concept, however if we take out the outliers the estimation suggest that the amount spent in this concept is only 26 mexican pesos. That represents a reduction of 87% when the data does not contain outliers.

Another way to study the effect of the outliers is to compute medians instead of means. Table 7.3 presents the estimated expenditure per concept using the median. Because of the high presence of zeros in the data, we compute two estimates, the first one with the complete data and the second one taking out the zeros.

	FOOD	DRESS	HOUSE	MAINT	HEAL	TRANS	EDUC	PERS
All data	$1,814	$165	$1,278	$195	$0	$616	$121	$277
Without zeros	$1,828	$243	$1,282	$199	$95	$748	$434	$281

Table 7.3: Estimates of median expenditure per concept with and without zeros. The amounts are in mexican pesos.

The expenditure estimates using the median gave values closer to the estimation obtain without outliers in Table 7.2. These results show that the ENIGH 2010 has outliers and these atypical values are causing an over estimation of the average expenditure per concept.

The effect of the zeros on the estimates is clearly shown in Table 7.3 for the concepts health, dress & shoes and education & hobbies where the estimates increase

significantly when the zeros are not considered. Even though the proportion of zero values is not significantly large in every concept, only 8,599 cases have information on all 8 expenditure variables, this represents only 31% of the total. This suggests the need for a procedure to estimate the complete data.

The previous results suggest that the use of a method that produces estimates that are not sensitive to outliers and also estimates missing values will be an advantage in this case. For these reasons we proposed to use our robust lower rank approximation to analyse the expenditure behaviour.

7.3 Results of applying the model

In the robust approximation of a rank one matrix to this data we will produce two vectors. One of the vectors will be the estimated amount spent per person for each household. The other vector will be the estimated percentage that the population spend in each concept on average.

The ENIGH 2010 has 27,655 cases. As mentioned our method allows a certain level of incomplete data depending on the number of complete cells per row an column which cannot be more than half. This means that we need observations that have information different to zero in at least 5 concepts. From the total cases we have that 26,567 cases comply with this condition, this means that we will be working with 96% of the total observations.

Our estimation of the vector of the total amount spent per person at each household has a skewed distribution as can be seen in Figure 7.2. The mean is $6,858, the maximum value of $106,339 and only 44 observations are greater than $40,000.

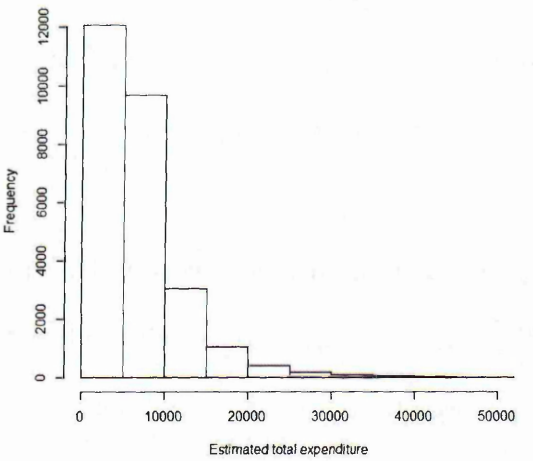


Figure 7.2: Histogram estimated total expenditure per person in each household.

Figure 7.3 presents the average distribution of the expenditure. These percentages are obtained considering all the households and they represent how much of the total amount is spent in each concept. We observed that more than 60% of the expenditure per household goes to food and house expenses. Also while 14% is allocated to transport and communications only 2% goes to health expenses.

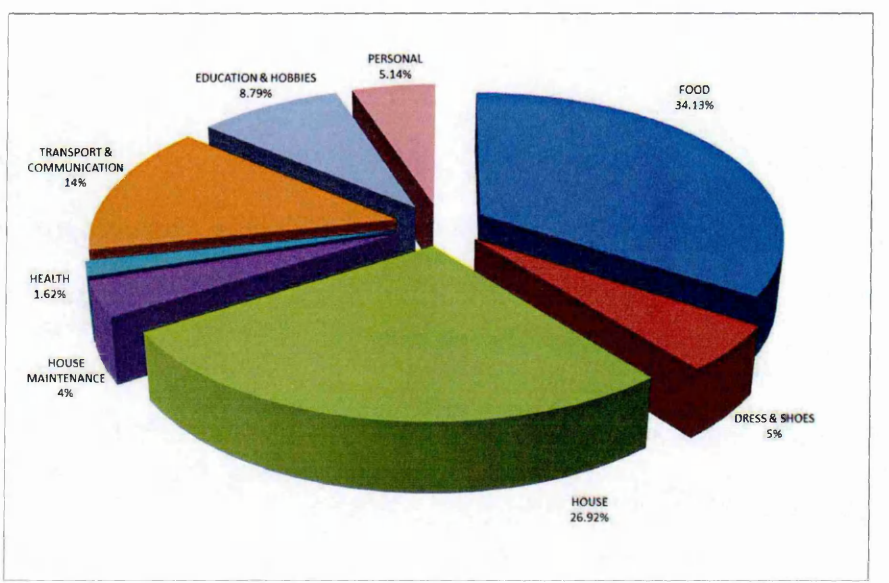


Figure 7.3: Estimated average expenditure distribution per person in each household.

These results can be compared with the official data published by INEGI (2011) and presented in Figure 7.4. The observed differences between concepts might been caused by the standardization we did of the households by the number of individuals in each house. So the amounts we use are average per person in each household and not total expenses of the household. This standardization gives a different comparison between households of different sizes.

Despite the difference of the information being by household or person per household, the distribution of the expenditure obtained with our method is similar to the published by INEGI in the concepts considered.

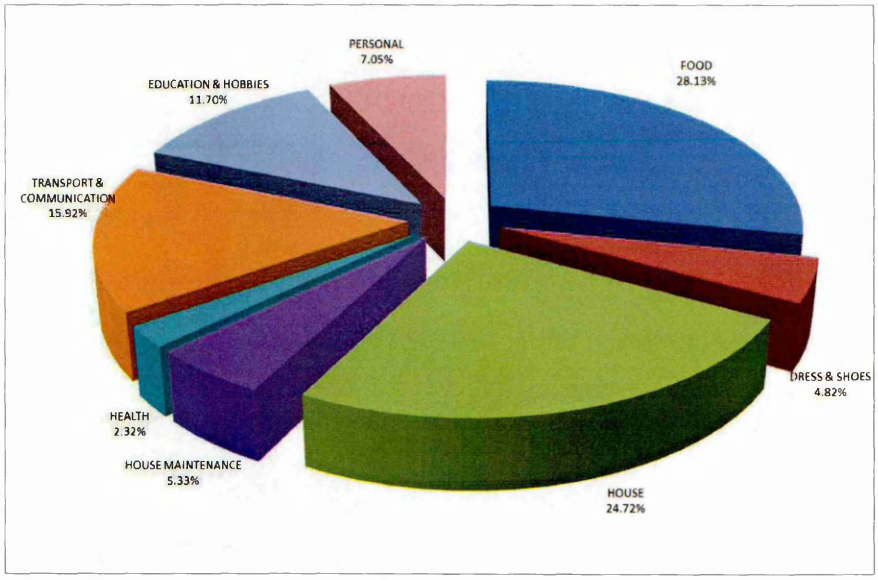


Figure 7.4: Distribution per household of the expenditure for each concept for 2008 and 2010.

Besides the two mentioned vectors, the model also produces a matrix of weights. These weights can be used to detect cases with possible errors in the data. The weights will be near to zero if the original amount differs from what will be expected on that concept for that particular household. So for cases where weights are low in one or two concepts it could mean that a mistake in those amounts has been made. In Figure 7.5 we present the weights for all the elements in our matrix. The

plot shows that for values greater than $31,000 (the horizontal line in the graph) the model is assigning weights equal to zero. This suggest that expenditures over 31,000 mexican pesos each trimester per individual in a household in any concept does not follow the general pattern of expenditure and those observations might be errors worth verifying.

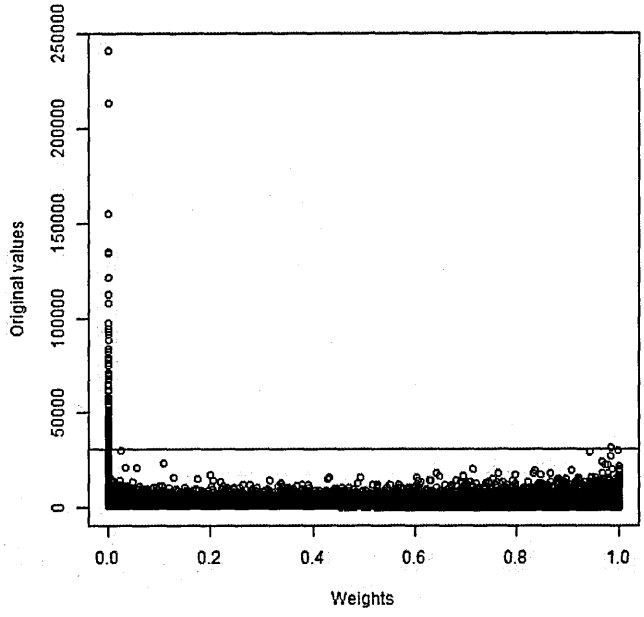


Figure 7.5: Weights assigned by our method vs original amount expended

Our method considers that if an observation has a weight smaller than 0.75 then this value should be marked as outlier. With this definition, 20.36% of the cases are outliers. But the weights are used to control the effect of the observations on the estimates. For example we could have an observation that the model assigned a weight of 0.74 and even though it is an outlier, the observation only counts as much as 0.74 of a non-outlier to estimate the total household expenditure amount. That is how the algorithm attempts to use as much information as possible from the real data.

There are households with a distribution of the expenses very different to the rest of the households; these will tend to show low weights in more than four concepts. Our

results show that 0.14% of the cases were identified as having only one good concept, and only 6.93% of the households had low weights in more than 3 concepts. This means the approximation with the two vectors is fairly consistent with the majority of the data.

Using the weights and the two vector estimates (**a** total expenditure per person at each household. **b** estimated average expenditure distribution) we obtain an expected matrix of expenditure. This matrix, unlike the original, will not have missing values nor outliers.

In the following Table 7.4 we observe the estimates of the average amounts spend in each concept. These estimates correspond to the expected value from our method, and the table compares them with the averages from the original data (that includes zeros and outliers).

	FOOD	DRESS	HOUSE	MAINT	HEAL	TRANS	EDUC	PERS
Original	$2,420	$370	$2,254	$483	$206	$1,218	$787	$514
Original without outliers	$1,967	$215	$1,470	$220	$26	$751	$258	$312
Median without zeros	$1,828	$243	$1,282	$199	$95	$748	$434	$281
Our estimates	$2,341	$346	$1,847	$272	$112	$985	$603	$353

Table 7.4: Estimates of the average expenditure per concept with the original data and with the estimates obtained from our method.

From Table 7.4 we observe that the concept with the largest difference is health, this was expected because this concept is the one with the largest proportion of outliers and missing values, 15% and 51% respectively. The table also shows that for food and dress & shoes the amount expended is very similar in both cases. Meanwhile the estimates obtained with our method are significantly lower for the rest of the concepts and this is due to the elimination of the large values. As discussed previ-

ously due to the existence of possible errors and missing values in the data, the use of the amount expended per household estimated by the model might be a more reliable source of information.

The other vector obtained from the method estimates the percentage that on average the population spent in each concept. We have compared these results with the proportions obtained with the original data. The results plotted in Figure 7.6 show a positive difference in the amount expended in the concepts of food and dress & shoes, our method is estimating greater percentages than the values obtained with the original data. Our method also estimates smaller percentages for the other six concepts. This differences might be related to the presence of outliers in the original data.

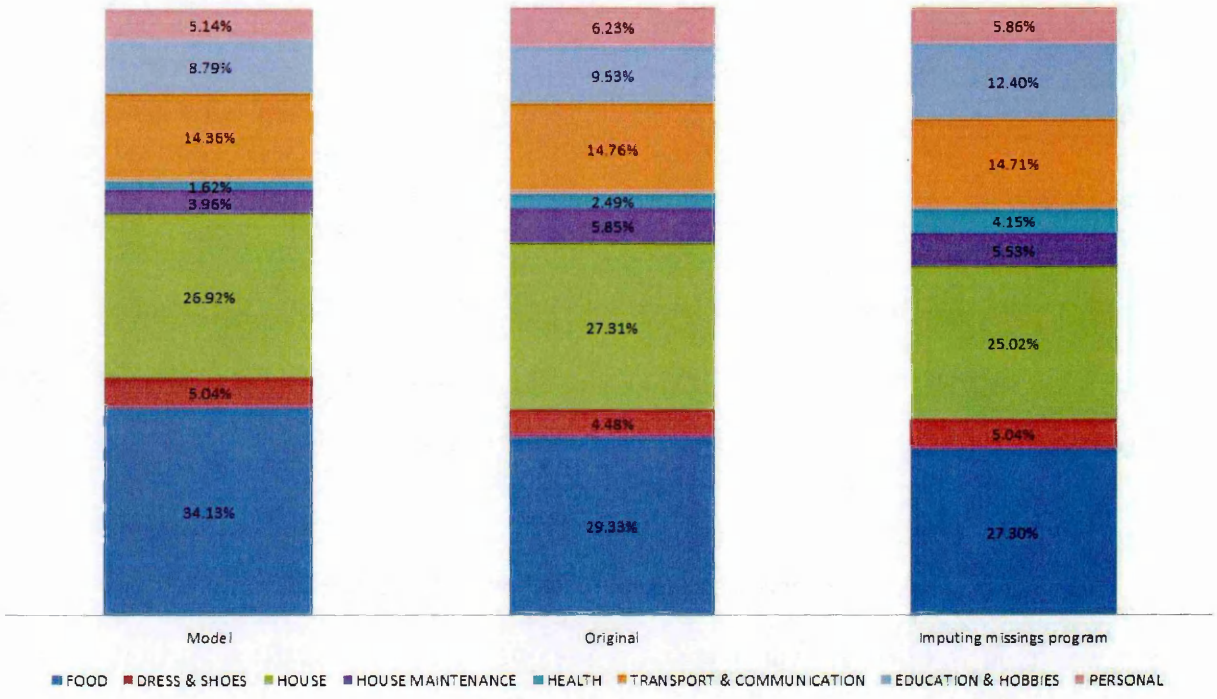


Figure 7.6: Distribution of the expenditure comparing the results of our model with the proportions obtained from the original data and from the results of applying a missing value imputation program.

For comparison purposes we apply the Multivariate Imputation by Chained Equations (MICE) algorithm programmed in R (Van Buuren, 2011). This method generates multiple imputations for incomplete multivariate data by Gibbs sampling, where the missing data can occur anywhere in the data. We have applied the algorithm only with the default options. The third column in Figure 7.6 presents the distribution of the expenditure calculated with the data set completed with the MICE algorithm. We observe that except for the health concept, the proportions obtained with the completed data set are similar to the proportions obtained with the original values. From these results is possible to say that this procedure could solve the problem of missing values however the effect of the outliers is still present.

In 2012 INEGI gathered the National Household Expenditure Survey (ENGASTO) and in October 2013 published the results (INEGI, 2012). We compared our results for the estimated expenditure by state with the results published by INEGI for the ENGASTO 2012. To make it comparable we multiply our estimates by the expansion factor and convert it to annual expenditure. The expansion factor is the value given in the survey to extrapolate the results into a national level. Figure 7.7 presents the average annual expenditure per person in each of the 32 Mexican states.

For ENGASTO 2012 the states with lowest annual expenditure per person were Tlaxcala with $29,659, Hidalgo with $29,459, Zacatecas with $28,669, Guerrero with $23,167 and Chiapas with $19,709. Coahuila, Baja California and Distrito Federal were the three states with the highest annual expenditure per person, $47,930, $51,532 and $59,203 respectively. Comparing with our results in Figure 7.7 we can see that the states with the lower expenditure are similar. Also Distrito Federal is the one with highest expenditure per person, and the difference in the amount with

the next state down is significantly large in both cases. However our estimate of the expenditure per person in Coahuila is not one of the highest values.

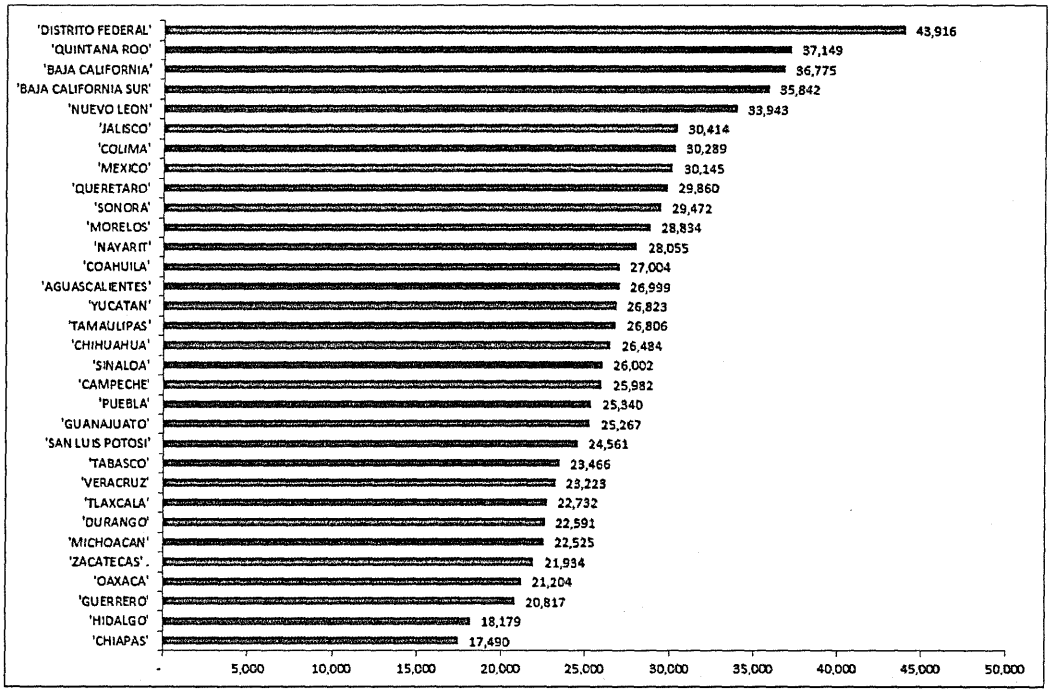


Figure 7.7: Estimated expenditure per year by state

Figure 7.7 show that average annual expenditure per person can go from 17,490 to 43,916 mexican pesos depending on the place of residence. Other variables such as number of people living in the house (home size) and the relationship between them (home type), the socio economic level and the locality size are expected to have an influence in this average expenditure per person.

We computed the correlation between the estimated expenditure and these variables. The results show that the strongest correlation is with the number of people living in the house (0.43), then socio economic level and locality size with 0.34 and 0.35 respectively. Analysing the correlation between these variables; we observed that the locality size and the socio economic level have a 0.70 correlation coefficient, meanwhile the home size and the socio economic level have a 0.15 correlation coefficient . Due to this correlation between variables we will only analyse the average

expenditure per person by home type and socio economic level.

SOCIO ECONOMIC		HOME TYPE	
very low	\$2,922	**unipersonal**	\$13,012
low	\$4,150	**nuclear**	\$6,541
medium	\$5,299	**nuclear+relatives**	\$5,258
high	\$6,196	**nuclear+non relatives**	\$5,399
very high	\$8,338	**co residents**	\$12,178

Table 7.5: Estimates of the average individual expenditure by socio economical and home type variables.

Table 7.5 shows that people from a very low socio economic level spend on average 2,922 mexican pesos every three months while a person from a very high socio economic level spends almost three times more. For households with only one resident the expected expenditure is 13,012 mexican pesos not very different to households of co residents. However for family households (parents and children) the average expenditure per person drops more than half. These results, as well as the plot for the annual expenditure per state, suggest that the average expenditure variates depending on the characteristics of the household.

We studied if this difference is also present in the distribution of the amount spent across the eight concepts. We know people from low socio economic levels spend less but do they spend it in a different way than households from other levels? So in the following figure we have the plot of the expenditure distribution for each socio economic level.

Figure 7.8 shows that as the socio economic level decreases the proportion of money spent on food increases. The contrary happens to the proportion spent in housing, higher the socio economic level higher the proportion spent on housing. The rest of the concepts remain more or less stable across the socio economic variable. These results are comparable with the data published for income by INEGI for the ENIGH

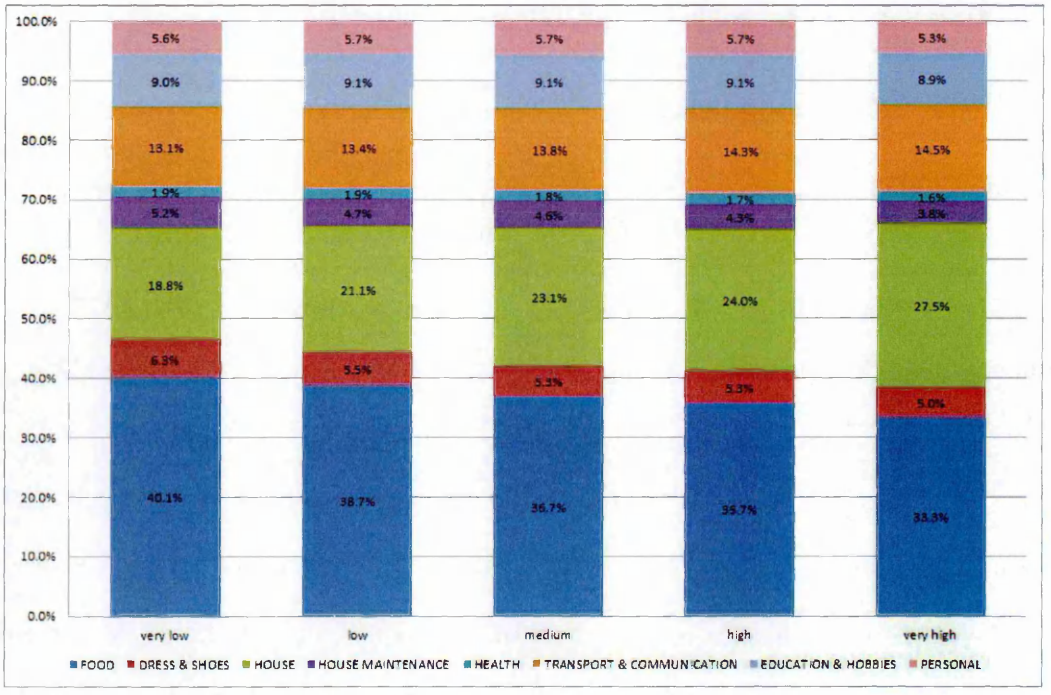
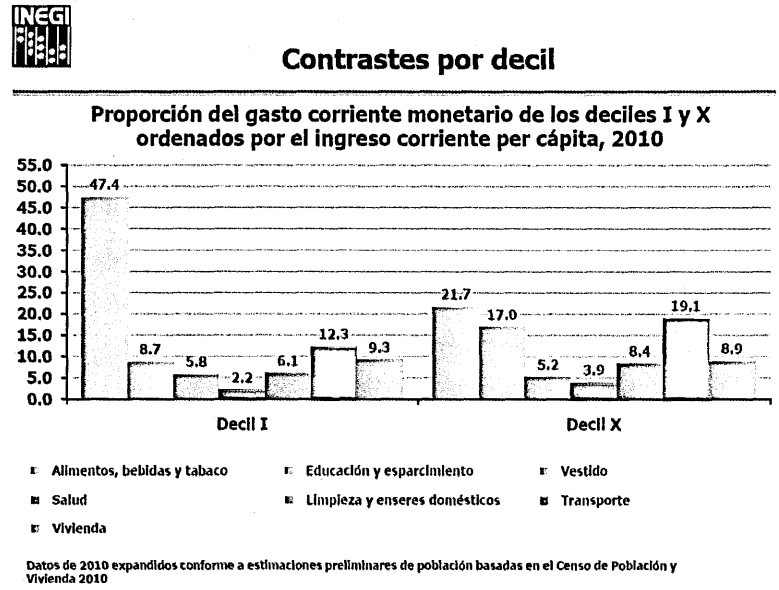


Figure 7.8: Distribution of the expenditure by socio economic levels

2010 (INEGI, 2011).

Figure 7.9 presents the graph of the expenditure distribution for first and tenth decile of income. To corroborate that our estimates are consistent, we will expect to have a similar tendency of the expenditure distribution between very low socio economic level and the first decile of income according to ENIGH 2010, and a similar relation between very high socio economic level and the last decile of income.

Both Figures 7.8 and 7.9 shows a relationship between low income and high percentage of expenditure on food and high income related to higher expenditure in transport & communications. However Figure 7.8 does not show the relation between high income and higher expenditure in education & hobbies, a relation that is observed in Figure 7.9. It is possible that this difference is caused by the fact that we are comparing socio economic level in Figure 7.8 and income in Figure 7.9, but in the publications found the results were not presented by socio economic level.



FOOD=Alimentos, bebidas y tabaco
EDUCATION & HOBBIES= Educacion y esparcimiento
DRESS & SHOES=Vestido
HEALTH=Salud
HOUSE MAINTENANCE=Limpeza y enseres domesticos
TRANSPORT & COMM= Transporte
HOUSE=Vivienda

Figure 7.9: Distribution of the expenditure by income

Considering all the comparisons made in this section between the estimates obtained with our method and the different results published by INEGI, it is possible to conclude that the average expenditure per person and the distribution of the expenditure in the eight concepts estimated by our method seem to be fairly congruent with the analysis already done and published by INEGI.

7.4 Sensitivity analysis

In order to study the method's performance we analysed the effect the data set composition has in the estimates of vectors related to the expenditure distribution. The hypothesis is that the method is robust enough to give reliable estimates independently of the data set composition. To do this analysis we selected randomly

households from the data base from the 26,567 households that have non-zero information in at least 5 expenditure concepts. This selection becomes a subset to which the method is applied and the vector of expenditure distribution is estimated.

We set the size of the subset to be 1000 households and we repeat the process for 1000 subsets. The results of the 1000 estimated expenditure distributions by concept are plotted in the histograms in Figure 7.10.

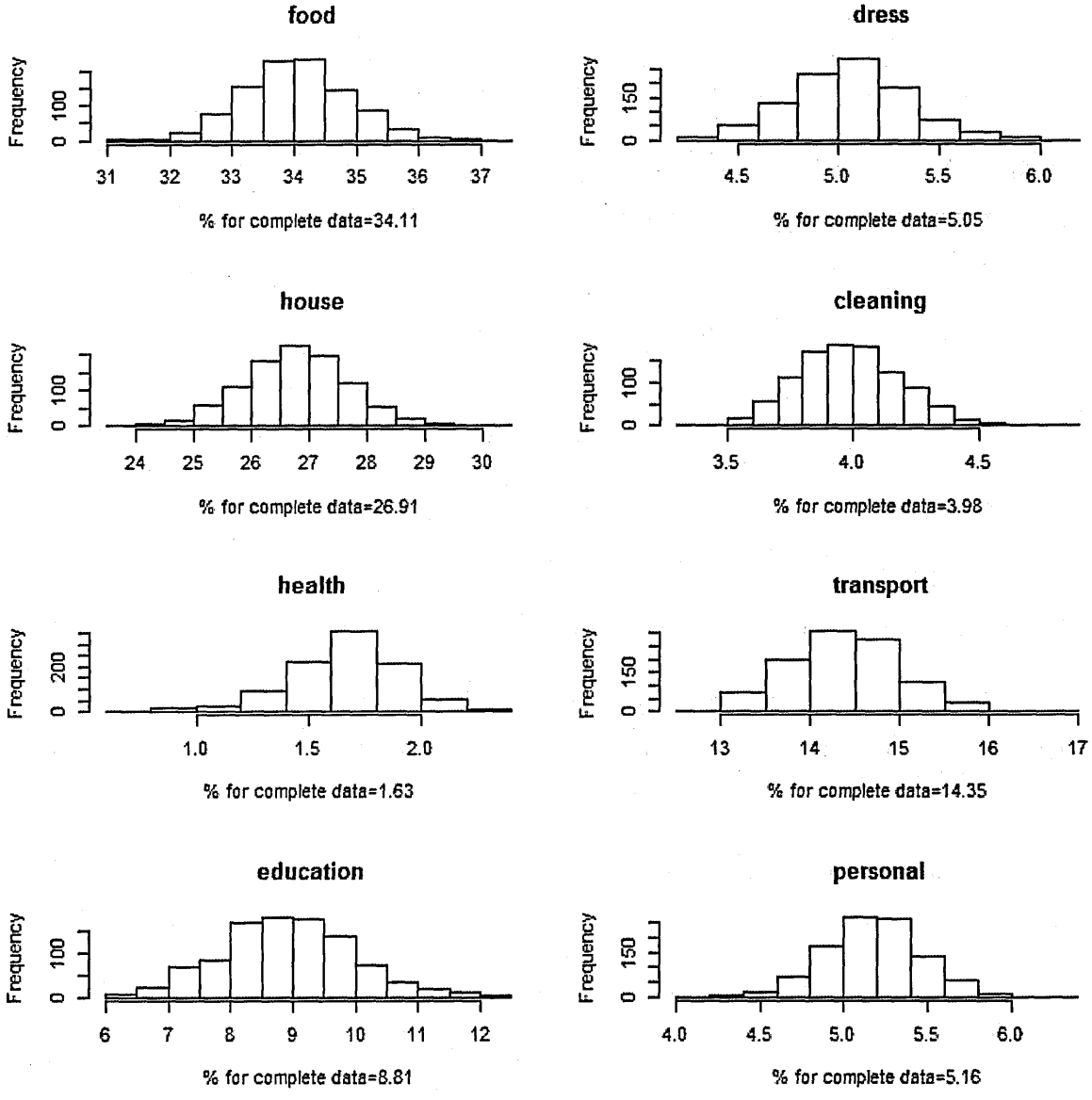


Figure 7.10: Histogram of the households distribution expenditure

From the histograms we have that for example for the food concept, each of the 1000 sub groups of household spend between 31 to 37 percent in this concept, with the mean around 34%. Figure 7.10 also has the information about the percentage of expenditure estimated for the complete ENIGH 2010 data base, in the case of food this percentage was 34.11%. The percentage of expenditure estimated using the complete data is always near the mean of the histograms for all the concepts. This suggests that our method is robust enough to give reliable estimates independently of the data set composition.

We conclude that the use of a method that produces estimates not sensitive to outliers and that at the same time estimates missing values is an appropriate procedure to analyse the households' expenditure behaviour. The advantage of estimating the missing values is that the usable information increases; so far the results published by INEGI use the raw data, not distinguishing between zero values because the amount spent was zero or because no information was provided. We have also shown the existence of large and influential outliers in the data, and our method proposed a way to treat them without losing all the information. Finally the results of average expenditure per person and the distribution of the expenditure in the eight concepts seem to be fairly congruent with other results published. These suggest that the use of our lower rank approximation algorithm for data with asymmetric outliers produces reliable results in this application.

Chapter 8

Conclusions, discussion and further research

This thesis presents the development of a method to measure the performance of competitors in an orienteering course. This method allows the comparison of an orienteer's performance with other competitors running the same course, and with his/her performance in different courses. The method proposed calculates a fitness measure and a navigational measure, analysing the effect of these two abilities separately on the overall performance. This differentiation between physical and navigational skills is an original approach to the orienteering performance analysis, and it allows a clear identification of areas that need improvement for a better course performance.

The approach used to construct these performance measures was through robust asymmetric lower rank approximation. In Chapter 3 we presented an algorithm that performs this asymmetric lower rank approximation through an iterative process. In Chapter 4 we tested with simulated data sets the sensitivity of the algorithm. We found that the algorithm performs well for data with asymmetric outliers, performs

similarly to standard algorithms when the data does not have outliers and performs poorly when the data has symmetric outliers.

Because of the nature of the data, this research focused on the reduction of a matrix $\mathbf{T}$ of dimension $n \times m$ with outliers into the product of two vectors $\mathbf{a}$ and $\mathbf{b}$. We discussed that the main reason for choosing a two vectors approximation was the easy interpretation of the results, with one vector relating to an orienteer's speed and the other to the distances between check points. In Chapter 4 we explored through simulations the option of improving the estimates by increasing the rank of the approximation to a rank equal to two. After comparing the results we concluded that the estimates obtained by the approximation with two vectors $\mathbf{a}$ and $\mathbf{b}$ produces estimates as good as the one obtained from approximating with matrices of the form $\mathbf{A}_{n\times 2}$ and $\mathbf{B}_{2\times m}$. In Chapter 5 we applied our algorithm to real orienteering data. The results show that an orienteer's speed and the distances between check points estimated through the matrix reduction produce reliable results. We believe approximations with ranks greater than two would not improve the estimates significantly, and because of the good results obtained by the two vectors approximation, the exploration of higher ranks is left for future research.

From the two vectors obtained with our robust and asymmetric lower rank approximation algorithm we defined two fitness measures and three navigational measures. The analysis done with orienteering data shows that an orienteer's performance measure proposed in Chapter 5 produces good estimates of both fitness and navigational abilities. The orienteering problem is complex in terms of the data because none of the courses are run more than once by each orienteer. The information available is the times between check points and the designed course distance. Every course

is different and the identification of specific orienteers across courses is not easy. Then an orienteer's performance measures proposed in this thesis can be used as a point of reference for new improved measures. Improvements on the fitness and navigational measures could be made if the identification of orienteers over different courses is done and the procedure allows the comparison over different events. Another approach will be to select a method that is not course dependent and for which more than one course is used to estimate an orienteer's performance. An orienteer's name could be used to identify orienteers in a course, because they are consistent and almost unique, however the mix of orienteers obtained in a particular course is not consistent. This means that to be able to use the orienteers as unit of analysis, we will need to have a very large amount of courses or knowledge of the courses where the orienteer participated. However because of the time needed to collect the data using the orienteer as unit, we decided to work with the course as unit. So the mentioned improvements are out of the scope of this thesis.

This research focuses on two aspects of orienteering, on one side the orienteer and the analysis of his/her performance has been mentioned in the previous paragraphs, on the other side is the course. As part of the course analysis we tried to relate the times to complete the course with some covariates such as number of controls, distance and climbing metres of the course. The results show that the correlation between the times and these variables is small and insignificant. We have explained this low correlation as a consequence of the fact that course designers use those variables to design a course of a given level of difficulty. The limitation on the information available about the characteristics of the courses stopped us going deeper into finding other covariates.

Because we could not use covariates to characterise the course, we presented in Chapter 6 a proposal to calculate the course technical difficulty based on the performance of the orienteers running that particular course. The results show that this approach produces a reliable technical difficulty measure that can be use to analyse differences between courses. We applied this difficulty measure to three different sets of colour coded courses. These types of courses have the property that by design the same colour courses should have similar technical level independently of the event. We constructed an empirical difficulty distribution for each of the three colour courses studied. The distribution was based on the times of 100 courses that took place in the UK between January 2013 and May 2014. This difficulty distribution was used to identify easy and hard courses over the difficulty range for each colour.

A different approach to analyse the course difficulty is through the ranking points. The ranking points awarded at each course depend on the time taken by all the ranked orienteers and the quality of those runners (which is measured by their points just before the event), so this information could be used to measure the course difficulty. However the ranking lists are rapidly and frequently updated so it was hard to obtain the ranking points of a large number of orienteers before each event. For this reason we did not use the ranking points in our analysis, but we think that if the data are collected further research could be done.

Besides an orienteer's performance measure and the course technical difficulty measure, this thesis also presents an innovative robust and asymmetric lower rank approximation algorithm. Even though this algorithm was constructed based in the orienteering problem, it can be used in cases where the data we want to analyse can be arranged as a matrix with asymmetric outliers and with certain dependency be-

tween the columns. An example of another application was presented in Chapter 7 where we show how this algorithm is applied to household expenditure estimates with good results.

The algorithm proposed in this thesis was based on the algorithm proposed by Maronna and Yohai (2008), with two original contributions; the definition of an asymmetric objective function, and the use of a robust scale parameter estimator that is updated at each iteration. In this matter further research can be done in studying the effect of different objective functions depending of the data analysed. The proposed modification to the scale parameter had an effect on the algorithm performance. First we observed that the estimates of our algorithm do not depend on initial values, so there is no need to set initial values to start the iterative process. Second the vector estimates seem to be closer to the real values. Finally, the value reached by the loss function is most of the time better than the point reached without the modification. This despite the fact that the updating modification had a direct implication in the convergence of the iterative process, because the algorithm does not converge to the minimum value found in the iterations, it converge to a slightly larger value.

The improvement mentioned in this iterative algorithm, gained by using different scale parameters and updating them, is an effect that might be possible to replicate in other iterative algorithms. It is sensible to think that other procedures with iterative linear regressions might respond as well as this algorithm when the update is introduced. It will also be worth exploring any iterative procedure whose unknown parameters are estimated only with initial values. Unfortunately that study is outside the scope of this thesis and the thoughts are left for future research.

In summary this thesis presents a robust and asymmetric lower rank approximation algorithm to reduce matrices with asymmetric outliers. It introduces an novel procedure to estimate orienteer's performance by differentiating between fitness and navigational abilities, and also propose a course difficulty distribution.

References

Allende, H., Frery, A., Galbiati, J., and Pizarro, L. (2006). M-estimator with asymmetric influence function: the g_A^0 distribution case. *Journal of Statistical Computation and Simulation*, 76(11):941–956.

Attackpoint (2004). Discussion: New split analysis. `http://www.attackpoint.org/discussionthread.jsp/message_3012`. [Online; accessed August 24, 2014].

Bird, S., Bailey, R., and Lewis, J. (1993). Heart rates during competitive orienteering. *British Journal of Sport Medicine*, 27:53–57.

Bird, S., Balmer, J., Olds, T., and Davison, R. (2001). Differences between the sexes and age-related changes in orienteering speed. *Journal of Sport Sciences*, 19:243–252.

BOF (2014). Appendix b: Course planning. `http://www.britishorienteering.org.uk/images/uploaded/downloads/events_appendix_b_2014.pdf`. [Online; accessed June 18, 2014].

Colin, C., Xuming, H., and Ying, W. (2008). Lower rank approximation of matrices based on fast and robust alternating regression. *Journal of Computational and Graphical Statistics*, 17(1):186–200.

Croux, C., Filzmoser, P., Pison, G., and Rousseeuw, P. (2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, 13:23–26.

Gabriel, K. and Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21:489–498.

Gjerset, A., Johansen, E., and Moser, T. (1997). Aerobic and anaerobic demands in short distance orienteering. *Scientific Journal of Orienteering*, 13:4–25.

Householder, A. and Young, G. (1938). Matrix approximation and latent roots. *American Mathematical Monthly*, 45:165–171.

Hubert, M., Rousseeuw, P. J., and Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, 23(1):92–119.

INEGI (2010). ENIGH 2010. `http://www.inegi.org.mx/est/contenidos/Proyectos/Encuestas/Hogares/regulares/Enigh/Enigh2010/tradicional/default.aspx`. [Online; accessed July 15, 2014].

INEGI (2011). Presentacion de resulatdos ENIGH 2010. `http://www3.inegi.org.mx/rnm/index.php/catalog/36/download/2122`. [Online; accessed July 15, 2014].

INEGI (2012). ENGASTO 2012. `http://www.inegi.org.mx/inegi/contenidos/espanol/prensa/Boletines/Boletin/Comunicados/Especiales/2013/octubre/comunica9.doc`. [Online; accessed July 15, 2014].

Jackson, J. E. (1991). *A user's guide to principal components.* John Wiley & Sons, Ltd.

Jensen, K., Franch, J., Krkkinen, O., and Madsen, K. (1994). Field measurements of oxygen uptake in elite orienteers during cross-country running using telemetry. *Scandinavian Journal of Science and Medicine in Sports*, 4:234–238.

Jensen, K., Johansen, L., and Krkkinen, O. (1999). Economy in track runners and orienteers during path and terrain running. *Journal of Sports Sciences*, 17:945–950.

Johansson, C., Lorentzon, R., Rasmuson, S., Reiz, S., Haggmark, S., Nyman, H., and Fugl-Meyer, A. (1988). Peak torque and OBLA running capacity in male orienteers. *Acta Physiologica Scandinavica*, 132:525–530.

Kolb, H., Sobotka, R., and Werner, R. (1987). A model of performance-determining components in orienteering. *Scientific Journal of Orienteering*, 3:71–81.

Korhonen, M., Mero, A., and Suominen, H. (2003). Age-related differences in 100-m sprint performance in male and female master runners. *Medicine & Science in Sports & Exercise*, 38(8):1419–1428.

Krzanowski, W. J. (1998). *Principles of Multivariate Analysis.* Oxford University Press.

Larsson, P., Burlin, L., Jakobsson, E., and Henriksson-Larsen, K. (2002). Analysis of performance in orienteering with treadmill tests and physiological field tests using a differential global positioning system. *Journal of Sport Sciences*, 20:529–535.

Lui, L., Hawkins, D. M., Ghosh, S., and Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100(23):13167–13172.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics.* John Wiley & Sons, Ltd.

Maronna, R. A. and Yohai, V. J. (2008). Robust low-rank approximation of data matrices with elementwise contamination. *Technometrics*, 50(3):295–304.

Moser, T., Gjerset, A., Johansen, E., and Vadder, L. (1995). Aerobic and anaerobic demands in orienteering. *Scientific Journal of Orienteering*, 11:3–30.

Naismith, W. (1892). Untitled. *Scottish Mountaineering Club Journal*, 2(3):135.

Ottosson, T. (1996). Cognition in orienteering: theoretical perspectives and methods of study. *Scientific Journal of Orienteering*, 12:66–72.

Peck, G. (1990). Measuring heart rate as an indicator of physiological strees in relation to orienteering performance. *Scientific Journal of Orienteering*, 6:26–42.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283.

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–651.

Scarf, P. (2007). Route choice in mountain navigation, Naismith's rule, and equivalence of distance and climb. *Journal of Sport Sciences*, 25:719–726.

Seiler, R. (1996). Cognitive process in orienteering a review. *Scientific Journal of Orienteering*, 12:50–65.

Troeng, M. (2008). Winsplits: Terms, concepts, and algorithms. `http://obasen.orientering.se/winsplits/help.aspx?topic=terms&lang=en`. [Online; accessed August 24, 2014].

Van Buuren, S., G.-O. K. (2011). mice: Multivariate imputation by chained equation in r. *Journal of Statistical Software*, 45(3):1–67.

Verboon, P. and Heiser, W. (1994). Resistant lower rank approximation of matrices

by iterative majorization. *Computational Statistics and Data Analysis*, 18:457–467.

Wang, F.-K. and Lee, C.-W. (2011). M-estimator with asymmetric influence function for estimating the Burr type iii parameters with outliers. *Computers and Mathematics with Applications*, 62:1896–1907.